PROTEIN EVOLUTION: MAPPING THE FITNESS LANDSCAPE AND THE ROLE OF CONSTRAINTS IMPOSED BY THE GENETIC CODE

by

Elad Firnberg

A dissertation submitted to The Johns Hopkins University in conformity

with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

October, 2013

ABSTRACT

Mutations are central to evolution, providing the genetic variation upon which selection acts. A mutation's impact on fitness can be positive, negative, or neutral. Knowledge of the distribution of fitness effects (DFE) of mutations is fundamental for understanding evolutionary dynamics, molecular-level genetic variation, complex genetic disease, the accumulation of deleterious mutations, the molecular clock, and the impact of constraints imposed by the genetic code. In order to facilitate study of the DFE, we developed a new mutagenesis technique, termed PFunkel, by which large gene libraries can be created in a single day, single tube reaction with user-control over the type and position of mutations as well as the number of mutations per gene. We used PFunkel to create several types of libraries of the *E. coli TEM-1* β-lactamase gene. By analyzing adaptive mutations in these libraries we found that the architecture of the genetic code significantly constrains the adaptive exploration of sequence space. However, the constraints endow the code with the ability to restrict access to amino acid mutations with a strong negative effect and, most remarkably, the ability to enrich for adaptive mutations. Furthermore, we present a comprehensive DFE for codon substitutions of the *TEM-1* gene and amino acid substitutions in the TEM-1 protein. This DFE provides insight into the origin of the genetic code, support for the hypothesis that mRNA stability dictates codon usage at the beginning of genes, an extensive framework for understanding protein mutational tolerance, and evidence that mutational effects on protein thermodynamic stability shape the DFE. Contrary to prevailing expectations, we find that deleterious effects of mutation primarily arise from a decrease in specific protein activity and not protein cellular levels.

Advisor: Professor Marc Ostermeier

Readers: Professor Michael Betenbaugh

Professor Kyle W. Cunningham

ACKNOWLEDGEMENTS

I would like to thank my advisor, Marc Ostermeier, for the opportunity to work in his laboratory. Marc has taught me a great deal about conducting scientific research and about the broader academic universe. My undergraduate training in traditional chemical engineering included very little biology. Therefore, the great majority of my knowledge and training in molecular biology and protein engineering has come thanks to Marc and my experience working in his lab. I especially appreciate his management style - both accommodating and always available to his students.

Johns Hopkins University and its community has been a wonderful place to live and work over the past five years. I have met some of my best friends and colleagues here at JHU and in the broader Baltimore community. Most importantly, I met my wife, Lisa, here in Baltimore. She has been a tremendous source of support and love, bringing her strong organizational and planning abilities to my life while accommodating my irregular research schedule.

I would like to thank all the past and current members of the lab for making it truly enjoyable to come into work every day. Special thanks goes to Barrett Steinberg and Nirav Shelat for being great friends both in and out of the lab. I owe a special thanks to three gentlemen who I count as some of my best friends who served as groomsmen at my wedding. Clay Wright, my roommate, head brewmaster, and fishing buddy has taught me more than I ever wanted to know about brewing delectable homebrews. David Broesch has been my musical counterpart, rock-climbing partner, and the other half of the novelty-comedy-folk-duo Tenacious PhD. Brian Chaikind has been a great friend and constant source of laughs and comic relief.

Finally, I would like to thank my parents, Anat and Duby, who early on instilled the importance of education for my sister, Maytal, and I. They are a strong reason I am getting my PhD and Maytal is in medical school. They have been there to support me at all times, especially during a difficult period of illness in high school and college. I also thank Maytal for her support and putting up with her obnoxious big brother all these years.

I have learned that life, like science, is not easy and abounds with challenges. The journey will present with one obstacle after another. The key is persistence and to view these challenges as opportunities to improve oneself - academically, physically, emotionally and spiritually. This may be one of the more important lessons that I hope to practice and impart on others.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 THE RELATIONSHIP BETWEEN DNA AND PROTEIN IS DICTATED BY THE GENETIC CODE

The central dogma of biology dictates that DNA, the molecular information carrier, is transcribed into RNA and then translated into protein via a base triplet to amino acid conversion key known as the genetic code. The remarkable universality of the genetic code across all forms of life, several minor variations notwithstanding, is a testament to the common evolutionary origin of life. Spontaneous mutations in DNA, primarily point mutations, provide the genetic variation upon which natural selection acts to drive Darwinian evolution. However, as illustrated in Table 1.1, the genetic code limits the mutational exploration of sequence space (1), since single base changes in codons can access only about six of the nineteen possible amino acid substitutions and simultaneous multiple-base changes in a codon are rare (2). Furthermore, the genetic code is biased towards conservative amino acid mutations (3). Since any arrangement of codon assignments would necessarily constrain access to amino acid substitutions, what explanations can be offered on the origin and current architecture of the natural genetic code?

1.2 THEORIES ON THE ORIGINS OF THE GENETIC CODE

After Watson and Crick famously reported the structure of DNA in 1953 (4), a variety of theories have arisen on the origin of the genetic code. In 1954, Gamow proposed a ‘key-and-lock’ theory, noting that there were twenty different shaped ‘holes’

formed by four adjacent nucleotides (dictated by three nucleotides) in the DNA chain, and speculated that each hole could be associated with a different amino acid (5). Subsequently, after the code was deciphered, more realistic stereochemical models have been proposed (6).

In 1965, Sonneborn and Woese proposed what has become known as the adaptive theory (3,7). Noting the high degree of order in the codon table, they argued that this could have been achieved through a stochastic evolutionary process independent of any specific DNA-amino acid physical interaction. The similarity among synonymous codons and those coding for chemically similar amino acids could be explained from an error minimization perspective. Early in evolutionary history, translation error rates were very high such that

Amino acid properties			WT amino acid	WT Codon	Amino acid substitutions accessible by point mutation	All Amino acid substitutions accessible by point mutation
hydrophobic	aromatic		W, Trp	TGG	C, G, L, R, S	C, G, L, R, S
			F, Phe	TTC	C, I, L, S, V, Y	C, I, L, S, V, Y
				TTT	C, I, L, S, V, Y	
			Y, Tyr	TAC	C, D, F, H, N, S	C, D, F, H, N, S
				TAT	C, D, F, H, N, S	
	non-polar aliphatic		P, Pro	CCA	A, L, Q, R, S, T	A, H, L, Q, R, S, T
				CCC	A, H, L, R, S, T	
				CCG	A, L, Q, R, S, T	
				CCT	A, H, L, R, S, T	
		start	M, Met	ATG	I, K, L, R, T, V	I, K, L, R, T, V
		small	I, Iso	ATA	K, L, M, R, T, V	F, K, L, M, N, R, S, T, V
				ATC	F, L, M, N, S, T, V	
				ATT	F, L, M, N, S, T, V	
			L, Leu	CTA	I, P, Q, R, V	F, H, I, M, P, Q, R, S, V, W
				TTA	F, I, S, V	
				CTC	F, H, I, P, R, V	
				CTG	M, P, Q, R, V	
				TTG	F, M, S, V, W	
				CTT	F, H, I, P, R, V	
			V, Val	GTA	A, E, G, I, L	A, D, E, F, G, I, L, M
				GTC	A, D, F, G, I, L	
				GTG	A, E, G, L, M	
				GTT	A, D, F, G, I, L	
hydrophilic		small	A, Ala	GCA	E, G, P, S, T, V	D, E, G, P, S, T, V
				GCC	D, G, P, S, T, V	
				GCG	E, G, P, S, T, V	
				GCT	D, G, P, S, T, V	
			G, Gly	GGA	A, E, R, V	A, C, D, E, R, S, V, W
				GGC	A, C, D, R, S, V	
				GGG	A, E, R, V, W	
				GGT	A, C, D, R, S, V	
	polar uncharged		C, Cys	TGC	F, G, R, S, W, Y	F, G, R, S, W, Y
				TGT	F, G, R, S, W, Y	
			S, Ser	TCA	A, L, P, T	A, C, F, G, I, L, N, P, R, T, W, Y
				TCC	A, C, F, P, T, Y	
				AGC	C, G, I, N, R, T	
				TCG	A, L, P, T, W	
				TCT	A, C, F, P, T, Y	
				AGT	C, G, I, N, R, T	
			T, Thr	ACA	A, I, K, P, R, S	A, I, K, M, N, P, R, S
				ACC	A, I, N, P, S	
				ACG	A, K, M, P, R, S	
				ACT	A, I, N, P, S	
			Q, Gln	CAA	E, H, K, L, P, R	E, H, K, L, P, R
				CAG	E, H, K, L, P, R	
			N, Asn	AAC	D, H, I, K, S, T, Y	D, H, I, K, S, T, Y
				AAT	D, H, I, K, S, T, Y	
	negatively charged, acidic		D, Asp	GAC	A, E, G, H, N, V, Y	A, E, G, H, N, V, Y
				GAT	A, E, G, H, N, V, Y	
			E, Glu	GAA	A, D, G, K, Q, V	A, D, G, K, Q, V
				GAG	A, D, G, K, Q, V	
	positively charged, basic		H, His	CAC	D, L, N, P, Q, R, Y	D, L, N, P, Q, R, Y
				CAT	D, L, N, P, Q, R, Y	
			R, Arg	AGA	G, I, K, S, T	C, G, H, I, K, L, M, P, Q, S, T, W
				CGA	G, L, P, Q	
				CGC	C, G, H, L, P, S	
				AGG	G, K, M, S, T, W	
				CGG	G, L, P, Q, W	
				CGT	C, G, H, L, P, S	
			K, Lys	AAA	E, I, N, Q, R, T	E, I, M, N, Q, R, T
				AAG	E, M, N, Q, R, T	
stop			*	TAA	E, K, L, Q, S, Y	C, E, G, K, L, Q, R, S, W, Y
				TGA	C, G, L, R, S, W	
				TAG	E, K, L, Q, S, W, Y	

Table 1.1. Architecture of the genetic code and the amino acid substitutions allowed by point mutation.

primitive cells would have produced collections of mutated and mostly non-functional proteins from a single gene. Woese showed that the error rate in the third position of the codon is 100-fold higher than the error rate of the first position and 10-fold higher than the second position, as measured in an in-vitro experiment modeling high translation error rates. As a means of reducing error, it was relatively easy for primitive cells to shuffle around codon-amino acid associations to increase robustness to translation mistakes, compared to the difficulty of evolving improved translation machinery. The 3rd codon position, prone to the highest error rate would have the highest degeneracy such that changes to it would have the least effect on the resulting amino acid. Such a code compared to ambiguous error-ridden codon assignments would have given primitive cells a distinct advantage by reducing translation errors and allowing them to evolve and produce better proteins (3,7). Though Woese focused on error minimization for the translation process, the degenerate code also provides a protective factor against mutations in DNA. In a stark change of direction, a year later Woese would go on to propose a stereochemical theory of the genetic code by which codon assignments were based on steric interactions between amino acids and oligonucleotide codons (referring here to mRNA rather than the just-discovered tRNA) (8).

In 1968, Crick proposed the frozen accident theory, opposing the theories put forward by Woese, particularly the stereochemical theory. The main hallmark of Crick's theory posits that the genetic code is fixed or frozen as we see it today because any changes to it would be lethal to the organism. Regarding the universality and common origin of the code, Crick offers a spectrum of possibilities. In the extreme form, codon allocation to amino acids was entirely a matter of chance while in a less extreme form,

Crick's theory resembles the adaptive theory with some nuanced differences. In Crick's imagining of the primitive code, few amino acids were initially present and almost all the base triplets could be read. Initial proteins were therefore quite crude and few in number. As new amino acids emerged, they were assigned codons similar to the codons of related amino acids so that the new amino acids, if substituted for, would not strongly disrupt the protein – thereby leading to an error-minimizing and conservative code. As proteins became more complex and numerous, a point was reached at which the code could no longer be changed without disrupting too many proteins, hence it became frozen (9). This led to the universality of the code today due to the extreme difficulty in changing it after it became fixed.

According to the co-evolution theory, proposed by Wong in 1975, the codon table evolved together with the biosynthetic pathways of emerging amino acids. Thereby, amino acid precursors of downstream amino acid products in the biosynthetic pathway would have codons that differ by only one base. For example, glutamine (codons CAA, CAG) is a biosynthetic precursor of histidine (codons CAU, CAC), and their respective codons differ by a single base. As the amino acid histidine emerged as a biosynthetic product of glutamine, the codon table was expanded to code for the new amino acid. However, Wong noted there are four amino acid precursor-product pairs that do not fit the theory and he suggests this may be due to changes in codons at a primordial stage (10).

These basic theories have been further developed primarily through theoretical and simulation approaches (6). In this work, we propose a new hypothesis that the code is arranged to facilitate the evolution of proteins by making adaptive mutations more likely

than under alternate codes. This would have given an adaptive advantage over other competing codes early in evolutionary history. In Chapters 3 and 4, we address this hypothesis experimentally by analyzing the distribution of adaptive mutations in three different genes. Chapter 5 discusses computational approaches that have been used to address these questions and how they may be applied to our theory.

1.3 SEQUENCE SPACE AND THE FITNESS LANDSCAPE

A difficulty long existed in Darwin's theory of evolution by natural selection in explaining the great diversity of functional protein sequences found in nature. The number of unique amino acid sequences possible from which nature must sift to find functional proteins is an astronomical number. For example, for an average length protein of 300 amino acids, there are 20^{300} possible unique amino acid sequences - many orders in magnitude greater than the total number of atoms in the observable universe (estimated at 10^{82}). If functional sequences are so rare compared to non-functional ones, it would seem impossible for evolution to find them.

Sewall Wright was one of the first to discuss this problem from a population

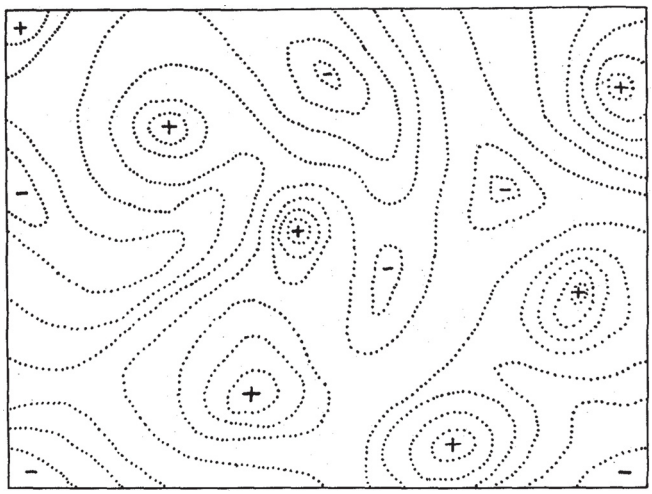


Figure 1.1. Wright's diagrammatic representation of a field of gene combinations in two dimensions instead of many thousands. Dotted lines represent contours with respect to adaptiveness (*11*).

genetics perspective. In his 1932 paper, as shown in Figure 1.1, Wright diagrammatically laid out the idea of a genetic mutational space as a field of gene combinations with contour lines representing adaptiveness – with natural selection forcing migration towards higher peaks (11). However, exactly how a species could travel from one adaptive peak to another amidst the vast number of inferior genetic combinations remained unclear.

In 1970, John Maynard Smith extended this work specifically to proteins and proposed a solution to this apparent contradiction. As natural selection fixes beneficial mutations, a protein diverges from its original sequence through a set of intermediates towards a new sequence providing improved function. Each mutational step must be functional or is eliminated. In this way, functional proteins form a continuous network that can be traversed by mutational steps (1). Thus, one can imagine a mutating protein sequence traversing a fitness terrain map whereby deleterious mutations form valleys, neutral mutations form flat plains, and beneficial mutations form peaks. On this fitness landscape, two sequences are neighboring if one can be converted to the other by a single amino-acid substitution. The physical environment influences the shape of this landscape, and therefore environmental changes lead to landscape changes, causing the protein sequence to travel up to new peaks. Maynard noted that in order for evolution by natural selection to be possible, there would have to be at least one beneficial accessible mutation for each step of the mutational trajectory. The concept of a protein space allowed Smith to more clearly formulate several important questions. For example, what fraction of potentially useful proteins are inaccessible (1)? The key to evolvability lies in the related

question - what fraction of accessible mutations are adaptive? Is the landscape smooth or rugged? Are there many local peaks, one global peak, or both (see Figure 1.2)?

1.4 THE DISTRIBUTION OF FITNESS EFFECTS (DFE) OF MUTATIONS

After Smith conceptualized the protein fitness landscape, many studies have sought to further characterize this framework theoretically and experimentally by quantifying the distribution of fitness effects (DFE) of mutations. Knowledge of the DFE is fundamental for understanding evolutionary dynamics, molecular-level genetic variation, complex genetic disease, the accumulation of deleterious mutations, the constraints imposed by the genetic code and the molecular clock.

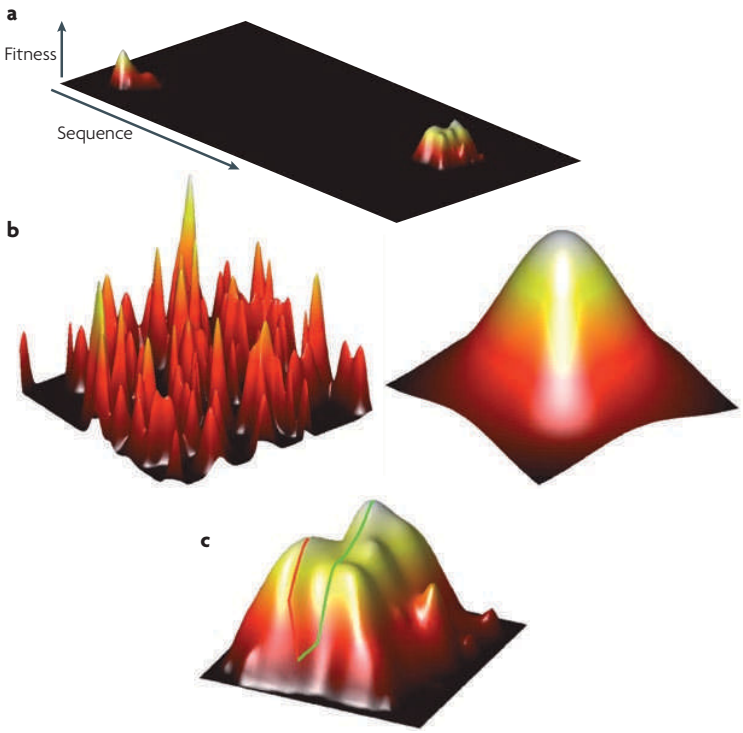


Figure 1.2. Theoretical renderings of fitness landscapes. The transition through black–red–orange–yellow represents increasing fitness. **a**. Although the details of this landscape are unknown, most sequences do not function (black) and the rare functional sequences are clustered near other functional sequences. This popular three-dimensional representation does a poor job of illustrating the numerous paths available to evolution and the numerous sequences in functional regions that do not encode functional proteins. **b.**

Evolution moves along networks of functional proteins that differ by a single amino acid, because selection requires a continuous uphill walk and does not permit the fixation of non-functional sequences. Epistasis occurs when the effect of one mutation depends on the presence of another, which can create landscape ruggedness and local optima. Landscapes could range from the rugged 'Badlands' landscape (left panel), which is nearly impossible to climb by mutational steps, to the 'Fujiyama' landscape (right panel), in which any beneficial mutation brings the search closer to the optimum. **c.** The presence of local optima might restrict some of the mutational paths uphill (red line). However, the large number of alternative routes leaves plenty of adaptive paths to a fitness optimum (green line) (12).

1.4.1 The DFE based on population genetics

Among more indirect methods, the population genetics approach makes use of DNA sequence data to infer the DFE of a population. In brief, this is done by comparing the number of nonsynonymous (or non-silent) mutations per total nonsynonymous sites, *dN,* to the number of synonymous (or silent) mutations per total synonymous sites, *dS,* for a particular effective population size, N_e. The model assumes that silent mutations are neutral and a fraction, *f*, of non-silent mutations are neutral (or deleterious but effectively neutral due to insufficient selective pressure) while the rest are deleterious. It can be shown that $dS = \mu$ and $dN = f\mu$, where μ is the nucleotide mutation rate. Therefore the ratio $dN/dS = f$ can be used to estimate the proportion of effectively neutral mutations and $1 - dN/dS$ estimates the proportion of deleterious mutations. If there are many adaptive mutations, *dN/dS* would overestimate the proportion of effectively neutral mutations (13). In hominids with N_e commonly 10,000 to 30,000, the ratio *dN/dS* is less than 0.3, suggesting fewer than 30% of nonsynonymous mutations are effectively neutral. In *Drosophila* and enteric bacteria with N_e in the millions and tens of millions, the *dN/dS* ratios suggest that at most 16% and 2.8%, respectively, are effectively neutral. Furthermore, these figures may be overestimated by 50% due to adaptive mutations. These observations indicate that the majority of amino-acid-changing mutations are deleterious (13).

1.4.2 Experimental small-scale approaches to characterize the DFE

Early experimental approaches for measuring the DFE have focused on characterizing random mutations introduced in a spontaneous or induced manner in various species such RNA viruses, bacteriophage, *E. coli*, *S. cerevisiae*, and *D. melanogaster*. The fitness distributions reported share similarities, with advantageous mutations accounting for 0–15% of mutations. In both RNA viruses and yeast nearly 40% of mutations are lethal. The DFE for deleterious mutations has appeared to be complex and multi-modal. For many datasets the DFE can often be fit to a gamma distribution (13).

Lind *et al.* studied the DFE of 126 mutations in two bacterial ribosomal proteins measuring growth rate as a proxy for fitness. It was found that most mutations produced a small fitness cost and the DFE for synonymous and nonsynonymous mutations was very similar with several low fitness nonsynonymous outliers. Fitness was found to correlate more with mutational effects on mRNA stability than specific amino acid changes (14). A limitation of this study is the small number of mutations characterized.

Peris *et al.* reported a DFE for 100 point mutations in the bacteriophage f1, a single-stranded DNA virus, and measured fitness as viral titer after infection of *E. coli* (15). The resulting DFE, as shown in Figure 1.3, was bimodal with ~20% of mutations lethal and

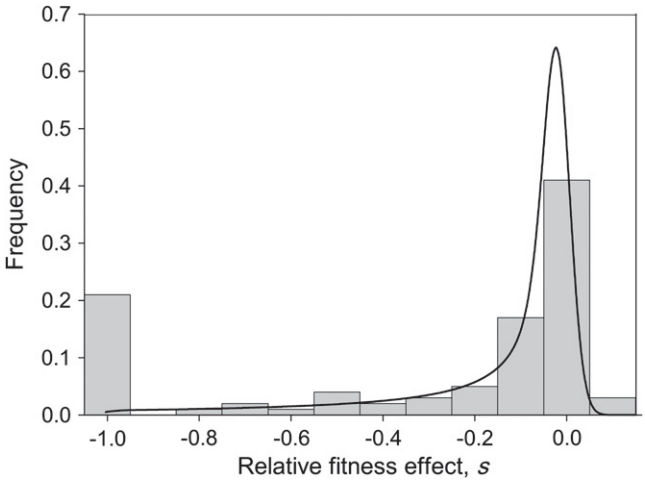


Figure 1.3. DFE of 100 point mutations in the bacteriophage f1. Non-lethal, non-beneficial mutations fit to a log-normal distribution (*15*).

viable ones reducing fitness by 11% on average and fitting a log-normal distribution. More than 90% of synonymous mutations were neutral while missense mutations reduced fitness by 37% on average. Interestingly, the DFE varied by the type of viral gene affected, with 47% of mutations lethal in replication-associated genes (genes II and X) while only 7% of mutations were lethal in an extrusion-associated gene (gene IV). Two beneficial mutations were identified in genes I and IV, involved in maturation and extrusion, respectively. Again, a limitation of this study is the small number of mutations characterized.

Wloch *et al.* reported a DFE for chromosomal mutations in the yeast *S. cerevisiae.* In this study, populations of yeast were propagated with mutational induction by ethyl methanesulfonate (EMS) treatment or inactivation of the mismatch repair machinery along with control populations. Following sporulation of diploid cells, the tetrad haploids were dissected and grown in individual colonies side-by-side on plates. Fitness was determined by growth rate inferred from measurement of colony size. Diploid cells that received a single mutation would manifest as two mutant colonies of equivalent size, and two wildtype colonies of equivalent size after tetrad dissection. The DFE of both mutagenesis methods was similar in the proportions of lethal mutations, ~42%, and about 30% lethal among spontaneous mutations in the control group. The EMS distribution showed notable bimodality. The average relative fitness of nonlethal mutations for both methods was similar, ~0.78, but lower than that for spontaneous mutations, ~0.9. While novel for presenting a DFE for random mutations in a eukaryotic genome, a limitation of this study is the questionable meaningfulness of a DFE covering such a small proportion

of genomic sequence space. Additionally, there is no characterization of the location or role of individual mutations, though such data would be very difficult to obtain.

1.4.3 Experimental large-scale approaches using deep sequencing

More recent studies have attempted a more systematic characterization of the DFE, taking advantage of advances in deep sequencing technologies. In 2010, Fowler *et al.* reported a high-resolution mapping of 600,000 variants of the human WW domain. The authors produced a phage-display library of the 56-residue domain in which each gene received an average of two DNA base mutations within the 99 base region (33 amino acids) that was targeted for mutagenesis. The library was subjected to six rounds of enrichment/selection for binding to the target peptide and then subject to Illumina deep sequencing, though the read-length limitation of this technology permitted sequencing of only 25 of the 33 mutated positions. From the data, they derived enrichment ratios for 405 of the 500 possible single amino acid mutations. The authors observed that 97.2% of the variants in the input library were depleted and therefore deemed deleterious (16). As might be expected, enrichment values were inversely related to folding energy as predicted by Rosetta. The authors also showed that a simple product model (not accounting for epistatic effects) could predict double mutant enrichment ratios from single mutant data with an r^2 correlation value of 0.68. In a subsequent publication two years later, Araya *et al.*, reanalyzed 47,000 variants from the same library by studying the relationship between single and double mutants in the context of protein stability. Returning to prediction of double mutant enrichment from single mutant data, they found that other models and better data could not improve prediction of double mutant scores beyond the simple product model – hence epistasis must be an intrinsic property of the

protein (see Figure 5.1). The authors then identified stabilizing mutations and activating mutations from the large dataset using an epistasis-based partner-potentiation metric which quantifies how much a single mutation improves the effect of its partner mutations. This method was able to identify stabilizing mutations, including several known ones, however some were neutral or destabilizing in subsequent measurements. This was attributed to the artificial phage display system, whereby the mutations may have stabilized the protein in the context of phage display but not in the protein alone (17). These types of artifacts are a limitation of DFE studies measuring protein fitness in a non-natural context such as phage display. A similar analysis was performed for activating mutations where it was found that both stabilizing and activating mutations could rescue deleterious mutations, but stabilizing mutations could rescue on average three times as many deleterious mutations as activating mutations.

A couple of recent studies have focused on the ubiquitination pathway. Starita *et al*. created a phage-display library of the C-terminal 102 amino acid portion of a murine E3 ligase, ubiquitination factor E4B (18). The library contained on average two nucleotide mutations per variant and 163,829 unique protein variants. The library underwent successive rounds of selection for in-vitro ubiquitin ligase activity using anti-Flag beads to enrich for Flag-tagged-ubiquitin-bound variants and then Illumina deep sequencing. Enrichment scores were determined for 932 variants with a single amino acid substitution of the 1,938 possible amino acid substitutions. Only 25 single mutants, 2.7% of the total exhibited 3-fold or greater enrichment over WT after three rounds of selection. Most of the activity-enhancing mutations were conservative and physically distant from the protein/protein interface. A limitation of this study is the artificial nature of the in-

vitro phage-display selection system used such that the protein is removed from its natural context.

A report by Roscoe *et al.* studied the DFE of most amino acid substitutions of the 76 amino acid yeast ubiquitin protein (19). After incorporating the mutagenesis library into the yeast host, cells were subjected to growth competition in liquid media for 50 hours. The relative abundance of mutants over time was determined by deep sequencing, providing a relative measure of fitness. The authors found that the great majority of solvent-accessible positions tolerated substitution of either greater than 16 amino acids or less than 5 amino acids. The majority of positions in the core tolerated 3-6 mutations to mainly hydrophobic amino acids. A major limitation of such growth competition experiments is that deleterious mutations are quickly depleted from the population and it is therefore difficult to determine an accurate and reproducible fitness measurement. This study could not reproducibly measure fitness for mutants with a fitness below ~40% of wildtype.

A recent publication by Whitehead *et al.* applied knowledge obtained from the DFE to improve two inhibitors of influenza hemagglutinin (20). These proteins were previously designed using computational methods and directed evolution. The authors used gene synthesis to perform NNK site saturation mutagenesis (each codon is randomized with all 32 combination of bases N=A,C,G,T and K=G,T) at ~50 amino acid positions for each of the two proteins. Using the yeast display format, the library was subjected to enrichment for binding to the influenza hemagluttinin target, and then the unselected and selected library was deep sequenced using the Illumina platform. After obtaining the DFE for the ~1000 amino acid substitutions in each of the proteins, shown

in Figure 1.4, the authors selected the ~10 most enriched mutations. They created secondary libraries of the two proteins randomized at these beneficial amino acid positions and repeated the enrichment

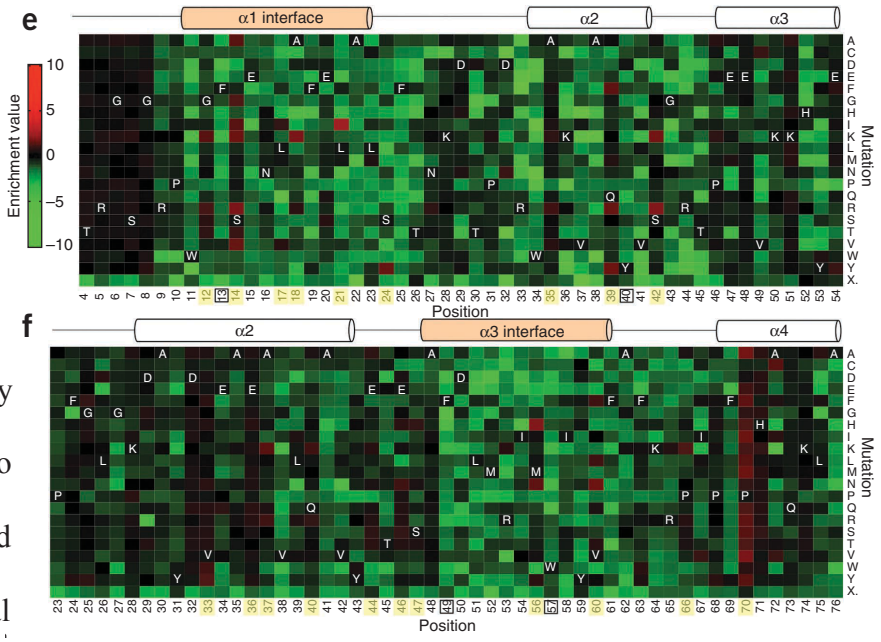


Figure 1.4. Heat map representing enrichment values for all possible single mutations in the two designed influenza inhibitor proteins. Most mutations are neutral or deleterious with little variation tolerated in the protein core or binding interface (20).

experiment. Analyzing the most highly enriched variants revealed an inhibitor from the first library with 8 amino acid mutations and a 28-fold lower binding constant, K_d, compared to the starting protein, and a variant from the second library with 5 amino acid mutations and a 25-fold lower K_d. These results demonstrate how the comprehensive landscapes can lead to greatly improved proteins. The beneficial combination of mutations discovered would have been unlikely to be found with an unbiased library of all combinations of 5 mutations since the diversity would be inaccessibly large. The traditional approach of directed evolution of iteratively selecting for the best variants and then combining their mutations would not have worked either since some of the mutations in the highest-affinity variant were not among the most enriched in the initial population. The results also illustrate how landscapes can be exploited to reprogram

interaction specificity of closely related targets by examining not only beneficial mutations but also neutral and deleterious ones.

1.4.4 Limitations of DFE studies

Each of the aforementioned DFE studies and others have advanced the field, however all suffer from certain limitations further elaborated in Chapter 4. The major limitations are: i. Only a small number of mutations are characterized, or the mutations are confined to one region of the protein (14,15,21). ii. The identity of the mutations is unknown (22). iii. The presence of multiple mutations per variant and interactions between them is unknown or unaccounted for (16,23). iv. An artificial selection system is used such that the protein is not functioning in its natural context (16-18) v. A growth competition experiment or experiments in which alleles are enriched based on threshold for function is used to measure fitness (16-20,23-25). Under these enrichment methods relative fitness measurements are highly dependent on the selective pressure used. Changing the selective pressure would significantly affect the relative fitness values measured. Furthermore, population size can affect the measured value of fitness due to stochastic effects. Finally, this method has limited ability to measure fitness for lower fitness variants due to their rapid depletion from the population. As detailed in Chapter 4, we overcame these limitations to provide the most comprehensive analysis to date of the distribution of fitness effects (DFE) for all codon substitutions of a natural gene and all amino acid substitutions of its corresponding protein.

1.5 THE DNA MUTAGENESIS TOOLBOX

In order to accomplish a comprehensive DFE study as mentioned above, a new mutagenesis method had to be developed in order to facilitate creation of large gene libraries with user-control over the type, location, and number of mutations per gene. No existing techniques, as summarized below, could produce such a library in an efficient manner.

Early attempts at mutagenesis involved using radiation or chemical mutagens to change DNA in a non-site-specific manner. In 1978, Hutchison *et al*. introduced in-vitro site-specific mutagenesis using a mutation-encoding oligonucleotide that annealed and primed strand-synthesis on a phage DNA template. The efficiency of the method yielded 15% mutants among the progeny phage (26).

1.5.1 Kunkel mutagenesis

In 1985, Kunkel introduced a method with improved efficiency (27). In Kunkel mutagenesis, a single-stranded uracil-containing phage DNA or phagemid template (containing the m13/f1 origin) is prepared by propagation in a *dut*$^{-}$*/ung*$^{-}$ strain of *E. coli*, which allows a certain amount of uracil incorporation into the DNA in place of thymine. The culture is then infected with m13 helper phage and the resulting phage particles are extracted and the single-stranded DNA purified from them. A 5' phosphorylated oligonucleotide complementary to the template, harboring the desired mutation, is annealed to the template at room temperature and then T4 DNA polymerase and T4 DNA ligase are used to complete the second strand and close the nick at 37 ^{o}C. Upon transformation of a normal *dut*$^{+}$ strain of *E. coli* with this product DNA, the cell's DNA repair mechanisms will select against and degrade the uracil-containing template strand,

giving preference to the mutated strand for propagation. Kunkel was able to achieve efficiencies of 50-70% mutated progeny (28), and others have since reported >90% efficiency in some reactions. The drawbacks of this method are the need to use a phagemid template and the need to create single-stranded DNA using phage. The method is best suited for making mutations at single sites though some have adapted it for making mutations at several sites simultaneously. As explained in detail in Chapter 2, Kunkel mutagenesis is not amenable for making libraries of variants with single mutations at different positions.

1.5.2 QuikChange site-directed mutagenesis

In 1996, scientists at Stratagene developed a whole plasmid site-directed mutagenesis method that eliminated the need for producing a single-stranded uracil-containing template (29). This method would later become popularly available as the QuikChange site-directed mutagenesis kit. Similar to a PCR, the method cycles a plasmid denaturing step, an annealing step for the mutagenic primers, and an extension step for the mutated strand. The use of a non-strand displacing polymerase, such as *Pfu*Turbo ensures that the polymerase is dislodged after going around the entire plasmid template, leaving a nick. In the final step, the template plasmid is destroyed by treatment with DpnI, a restriction endonuclease that recognizes and cleaves the sequence $GA^{m}TC$, since methylated adenine is present only in the template DNA purified from a dam^{+} strain of *E. coli.* An efficiency of at least 80% mutated progeny is reported for the method (30). Drawbacks of QuikChange are that the reaction outputs a low yield of nicked product DNA which results in very low transformation efficiency, and therefore it is not useful for making libraries.

1.5.3 Inverse PCR

Inverse PCR similarly uses complementary primers that encode the mutation, however unlike QuikChange the reaction results in exponential amplification of linear DNA (31). After PCR the ends are ligated together and the product transformed. The drawbacks of this method are that it depends on faithful amplification of the DNA termini and the low efficiency of blunt-ended ligations for reforming circular DNA. This method is also not suited for making complex libraries.

1.5.4 Combined-chain reaction

Combined-chain reaction uses a set of mutagenic primers and external primers to allow for multiple mutations on a linear template. Following PCR amplification, the linear ends are ligated and the product transformed. An efficiency of 87% was reported for introducing two simultaneous mutations (32). The drawbacks of this method are the need to design multiple sets of primers and the low efficiency of the ligation and sub-cloning step.

1.5.5 Error-prone PCR

Error-prone PCR is a commonly used method for generating genetic diversity. This method utilizes Taq polymerase and reaction conditions that reduce the fidelity of the polymerase. Conditions can be tuned to control the rate of mutation (33). Though it is a relatively simple and rapid method for creating mutations, this method has several drawbacks. The user lacks control over the type and location of mutations and there is strong mutational bias towards some types of substitutions over others. Furthermore, the method lacks the ability to effectively make most amino-acid substitutions, which require two or three base changes per codon.

1.5.6 Other random mutagenesis methods

A variety of other methods exist that rely on random DNA cleavage reagents or transposons for mutating short sequences of DNA (34-36). However these methods suffer from complex procedures, the inability to target the mutations, and other biases in the types of mutations created.

1.5.7 Limitations of mutagenesis techniques

All the methods described above are not capable of producing the type of library desired in this study. This type of library, henceforth termed a comprehensive codon mutagenesis (CCM) library, contains every possible codon substitution (replacement of each codon with the 63 other possible codons) in a coding sequence, such that each variant has only one codon substitution. Non-mutated (wildtype) genes, and genes with mutations in multiple codons are not desired in a CCM library. Each previously existing method has drawbacks preventing the ability to create a CCM library: i. Too few transformants are produced to create a comprehensive library. ii. Mutations can be targeted to only one position per reaction. iii. The mutational efficiency is too low, a significant fraction of progeny harboring the wildtype sequence. iv. One lacks control over the position and type of mutations created. v. One can only make point mutations, preventing access to most amino acid substitutions in the protein. The only ways to make a CCM library prior to the PFunkel method described in Chapter 2, was using site-directed mutagenesis (SDM) or gene synthesis. For SDM, one would have to perform a separate site-directed mutagenesis reaction for each codon position in the gene. For a 287 amino acid protein this would amount to 287 individual reactions, 287 transformations plated on 287 bacterial plates, and 287 recoveries of cells from plates. One can imagine

the extreme time, labor, and cost investment to make a single library in this manner. For gene synthesis, one would have to produce 287 individual constructs each 870 bases long, one for each mutated codon position of the gene. Then one would have to clone this library into a vector before placing in the host organism. Ordering such a library from a gene synthesis company would incur a cost close to $100,000, quite unfeasible for most researchers.

CHAPTER 2

PFUNKEL: EFFICIENT, EXPANSIVE, USER-DEFINED MUTAGENESIS

SUMMARY

We introduce PFunkel, a versatile method for extensive, researcher-defined DNA mutagenesis using a ssDNA or dsDNA template. Once the template DNA is prepared, the method can be completed in a single day in a single tube, and requires no intermediate DNA purification or sub-cloning. PFunkel can be used for site-directed mutagenesis at an efficiency approaching 100%. More importantly, PFunkel allows researchers the unparalleled ability to efficiently construct user-defined libraries. We demonstrate the creation of a library with site-saturation at four distal sites simultaneously at 70% efficiency. We also employ PFunkel to create a comprehensive codon mutagenesis library of the *TEM-1* ß-lactamase gene. We designed this library to contain 18,081 members, one for each possible codon substitution in the gene (287 positions in *TEM-1* x 63 possible codon substitutions). Deep sequencing revealed that ~97% of the designed single codon substitutions are present in the library. From such a library we identified 18 previously unreported adaptive mutations that each confer resistance to the ß-lactamase inhibitor tazobactam. Three of these mutations confer resistance equal to or higher than that of the most resistant reported *TEM-1* allele and have the potential to emerge clinically.

2.1 INTRODUCTION

An efficient and high-throughput mutagenesis strategy is an integral part of protein structure/function studies, directed evolution experiments for the discovery of novel proteins, and optimization of genetic elements in synthetic biology systems. Among the methods for in vitro mutagenesis, none offers a convenient, efficient and high-throughput approach for creating an extensive, user-defined library of variants in which single or multiple mutations can be located at any position. Site-directed mutagenesis methods such as Kunkel mutagenesis (27), QuikChange (30), and inverse PCR (31) are low-throughput methods. Combined chain reaction requires specially designed sets of primers and cloning of PCR products (32,37). Creating mutations by gene synthesis is comparatively expensive and requires sub-cloning of DNA. Error-prone PCR suffers from mutational bias, the inability to define the mutational composition, and the inability to effectively cause most amino acid substitutions, which require two or three mutations in a single codon. Methods that rely on random DNA cleavage reagents or transposons for mutating short sequences of DNA suffer from complex procedures and the inability to target the mutations (34-36).

Our method is inspired by Kunkel mutagenesis, a site-directed method that introduces mutations by using a mutation-encoding oligonucleotide (oligo) that anneals to a phage-derived, single-stranded uracil-containing circular DNA template. While the initial Kunkel protocol described making single base substitutions (27), other researchers have adapted the method for creating site-saturation libraries in a single region (38,39). The mutational efficiency of Kunkel mutagenesis is limited such that 50-90% of transformed colonies typically harbor the desired mutation while the remainder harbor

the wildtype sequence (28).

PFunkel, a conflation of *Pfu* DNA polymerase and Kunkel mutagenesis and pronounced "pee-funk-el", differs from Kunkel mutagenesis in a number of key ways that serve to increase the efficiency of the reaction and minimize the appearance of wildtype sequences in the resulting library. The major differences include (a) the use of a thermostable DNA polymerase and ligase, which enables a shift in the operating temperature of the reaction from 25-37 °C to 55-95 °C, (b) the option to use thermal cycling and stepwise addition of oligos to tailor the average number of mutations per gene, (c) the synthesis of a second mutated strand complementary to the first mutated strand that displaces the template strand, and (d) the in vitro degradation of the uracil-containing template and DNA products not in the desired covalently closed circular (cccDNA) form by the addition of uracil DNA glycosylase (UDG) and exonuclease III (Exo III). Additionally, we have developed a version of PFunkel that can be performed on any dsDNA plasmid template and avoids the use of phage.

We demonstrate PFunkel on the *TEM-1* gene encoding TEM-1 ß-lactamase by performing three types of mutagenesis experiments: (a) site-directed mutagenesis with 100% efficiency, (b) multiple-site mutagenesis, in which we create site-saturation libraries at four distal codons in *TEM-1* at ~70% efficiency, and (c) a new type of mutagenesis library called comprehensive codon mutagenesis. A comprehensive codon mutagenesis library consists of every possible codon substitution in the gene with only one codon substitution per library member (i.e. library members containing more than one codon mutated are not desired). Such a library is the equivalent of creating a site-saturation mutagenesis library at all positions in the gene. The degeneracy of this library

for *TEM-1* is 18,081 (287 codons x 63 possible codon substitutions). Deep sequencing of our library revealed that up to 97% of the possible 18,081 possible desired codon substitutions exist in the library and that the fraction of wildtype and variants with two or more codon substitutions in the library was ~13% and <3%, respectively.

2.2 MATERIALS AND METHODS

All enzymes were obtained from New England Biolabs (NEB) except PfuTurbo Cx hotstart DNA polymerase, which was obtained from Agilent Technologies. E. coli strain CJ236 and NEB 5-alpha F'Iq competent cells were obtained from NEB and strain DH5α was obtained from Invitrogen. R408 helper phage was obtained from Promega. All oligonucleotides were ordered from Integrated DNA Technologies. For the construction of library CCM-1 (machine-mixed degenerate oligos), oligos were ordered in 96-well format at the 10 nmole synthesis scale such that each oligo was provided at a concentration of 100 μM in DI water. For the construction of library CCM-2 (hand–mixed degenerate oligos), oligos were ordered in 96-well format at the 100 nmole synthesis scale such that each oligo was provided at a concentration of 100 μM in DI water. The secondary oligo P320, P-gcagaaattcgaaagcaaattcgac, was ordered with 5' phosphorylation. All other chemical reagents were obtained from Sigma-Aldrich.

2.2.1 Preparation of CJ236 competent cells

E. coli strain CJ236 was plated on LB-agar plates with 15 μg/mL chloramphenicol (Cm), 125 μg/mL deoxythymidine (dThd) and grown at 30 ^{o}C. Although it is usual to proceed with competent cell preparation from a single colony (especially if using a new cell stock validated by the manufacturer), we chose to first

confirm the desired strain phenotype, an optional step. A colony with the proper temperature- sensitive *dut-1* phenotype was identified by replica plating on M9 minimal media agar (40) supplemented with and without 125 μg/mL dThd and incubated at 30 °C and 42 °C for ~40 hours. A colony was selected which displayed the desired phenotype of stunted growth at 42 °C, which was improved with dThd (Personal correspondence with B. Weiss). This colony was used to prepare chemically competent cells (41). To prevent genetic drift and reversal of the *dut-1 ung-1* phenotype it is best to propagate CJ236 at ≤30 °C in dThd supplemented media. These conditions reduce uracil incorporation in DNA (an unfavorable mutagenic event leading to reversions of this phenotype) since uracil incorporation is unnecessary when propagating the strain. However, during preparation of uracil-containing ssDNA or dsDNA template, the strain should be grown at 37 °C without dThd for increased uracil incorporation.

2.2.2 Preparation of uracil-containing ssDNA template

pSkunk3-BLA is a 4.4 kB phagemid derived from pDIM-C8-BLA (42) in which the coding sequence of the Cm resistance gene was replaced with the streptomycin/spectinomycin (Sm/Spec) resistance gene. This phagemid was used to transform CJ236 cells which were then plated on LB-agar with 50 μg/mL Spec, 15 μg/mL Cm, and 125 μg/mL dThd and incubated overnight at 30 °C. A single colony was used to inoculate 10 ml of LB supplemented with Cm, Spec, and dThd as above, which was incubated with shaking at 30 °C overnight. The cell density of the culture was determined from the OD_{600nm} using the correlation 2×10^8 CFU/mL-OD_{600nm}. In a 20 ml test tube, 2 ml of TBG media (43) with 50 μg/mL Spec was inoculated with 2×10^7 CFU from the overnight culture and 1×10^8 pfu R408 helper phage for an MOI of 5. This

culture was incubated for 6 hours at 37 °C with shaking at 300 rpm. The culture was then centrifuged for 5 minutes at 16,100×g to pellet the cells, and the phage-containing supernatant recovered. Then 300 µL of 2.5 M NaCl/20% PEG was added to the supernatant and the mixture was incubated at 4 °C for 1 hour to precipitate the phage. The phage was pelleted by centrifugation at 20,817×g for 10 minutes at 4 °C. The liquid supernatant was discarded and the phage pellet resuspended in 150 µL PBS. The Qiagen QIAprep Spin M13 kit (#27704) was then used to purify ssDNA from the phage as per the manufacturer's directions. The absorbance at 260 nm of the ssDNA sample was measured using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific) and converted to a concentration using the relation 1.0 A_{260nm} = 33 ng/µL.

2.2.3 Site-directed PFunkel mutagenesis using a ssDNA template

All steps were performed in a pre-programmed Eppendorf Mastercycler personal thermocycler. A mutagenic oligo (5'-gacaccacgatgcatgcagcaatggc) encoding a c542a mutation in the *bla* gene was phosphorylated in a 50 µL reaction containing 1X T4 PNK buffer, 1 mM ATP, 5 mM DTT, 3.0 µM oligo and 10 units T4 PNK. The reaction was incubated at 37 °C for 1 hour, and the enzyme inactivated at 65 °C for 20 minutes.

The PFunkel reaction was prepared in a 0.5 mL eppendorf tube containing 1X PfuTurbo Cx hotstart DNA polymerase buffer, 10 mM DTT, 0.5 mM NAD+, 0.2 mM dNTPs, 1 µL of the kinase reaction, 1 µg (0.75 pmol) of dU-ssDNA template, 2.5 units PfuTurbo Cx hotstart DNA polymerase, and 200 cohesive end units Taq ligase in a total volume of 100 µL. The free Mg^{2+} concentration should be maintained between 0.5-2.5 mM since low concentration reduces polymerase fidelity while high concentration leads to nonspecific annealing of oligos (44). The volume of kinase reaction added should

therefore be minimized to maintain Mg^{2+} concentration in the mutagenesis reaction close to the 2 mM Mg^{2+} provided in the 1X polymerase buffer. The following denaturation/annealing/extension/ligation steps were performed: 95 °C for 3 min, 55 °C for 90 sec, 68 °C for 15 min and 45 °C for 15 min. Then 3.8 pmol of oligo P320 (5'-P-gcagaaattcgaaagcaaattcgac) was added and one more cycle of 95 ^{o}C for 30 sec, 55 °C for 45 sec, 68 °C for 10 min and 45 °C for 15 min was performed. Then 10 units of UDG and 30 units of ExoIII were added and incubated at 37 ^{o}C for 1 hr followed by an inactivation step at 70 °C for 20 min.

Five µL of the unpurified reaction was used to directly transform 100 µL of DH5α chemically competent cells (41). The entire transformation was plated on an LB- agar plate with 50 µg/mL Spec and incubated overnight at 37 °C. To obtain more transformants, the remaining DNA was purified using the Zymo DNA Clean & Concentrator kit according to the manufacturer's instructions and eluted in 15 µL of 1X EB. One uL was electroporated into 50 µL DH5α electrocompetent cells and then incubated with SOC recovery media for 1 hr at 37 °C with shaking at 250 rpm. The transformation was plated on LB-agar with 50 µg/mL Spec and incubated overnight at 37 °C. For the experiments of Table 2.3, the reaction was scaled down to 200 ng template and 20 µl volume.

2.2.4 Multi-site PFunkel mutagenesis using a ssDNA template

Four oligos were designed to introduce NNN random bases at codon positions 42, 104, 182, and 238 in the *bla* gene with respective sequences: 5'-gatcagttgggtnnncgagtgggttac, 5'-gaatgacttggttnnntactcaccagtcac, 5'-cgtgacaccacgnnncctgcagcaatg, 5'-aaatctggagccnnngagcgtgggtct. These oligos were

combined in equimolar amounts and phosphorylated in a 50 µL reaction containing 1X T4 PNK buffer, 1 mM ATP, 5 mM DTT, 6.0 µM total oligo and 10 units T4 PNK. The reaction was incubated at 37 °C for 1 hour, and the enzyme inactivated at 65 °C for 20 minutes.

The annealing reaction was prepared in a 0.5 mL eppendorf tube containing 1X PfuTurbo Cx hotstart DNA polymerase buffer, 2 µL of kinase reaction, and 1 µg of pSkunk3-bla ssDNA template in a total volume of 77 uL. The annealing was performed by heating to 95 °C for 3 min, then 55 °C for 10 min, and holding at 55 °C.

Meanwhile, in a separate PCR tube, 1X PfuTurbo Cx hotstart DNA polymerase buffer and 2.75 units of PfuTurbo Cx hotstart DNA polymerase were combined in a total volume of 5.5 µL. The hotstart polymerase was heat activated by heating to 95 °C for 3 min.

After the annealing step, 10 mM DTT, 0.5 mM NAD+, 0.2 mM dNTPs, 5 µL of the activated polymerase solution, and 200 cohesive end units Taq ligase was added bringing the total volume to 100 µL. The reaction was mixed by slowly and gently pipetting up and down. Extension and ligation of the mutant strand was performed at 65 °C for 15 min and 45 °C for 15 min. A total of 3.8 pmol of oligo P320 was added and one more cycle of 95°C for 30 sec, 55°C for 45 sec, 65°C for 10 min and 45°C for 15 min was performed. Five units of UDG and 2 units of ExoIII were added and the mixture was incubated at 37 °C for 30 min followed by an inactivation step at 70 °C for 20 min. The DNA was then purified using the Zymo DNA Clean & Concentrator kit according to the manufacturer's instructions and eluted in 15 µL of DI water. This volume was vacuum concentrated down to 1-2 µL, electroporated into 50 µL DH5α electrocompetent cells and

then incubated with SOC recovery media for 1 hr at 37 °C with shaking at 250 rpm. The entire volume was then plated on a Nalgene Bioassay dish (D4803; 245 mm × 245 mm × 25 mm) containing LB-agar with 50 μg/mL Spec and incubated overnight at 37 ^{o}C.

For the experiments of Table 2.3, the reaction was scaled down to 200 ng template and 20 μl volume.

2.2.5 Comprehensive codon mutagenesis by PFunkel using a ssDNA template

All steps were performed in a pre-programmed Eppendorf Mastercycler personal thermocycler. Equimolar amounts of 287 different mutagenic oligos were combined in a single tube at a total oligo concentration of 100 μM. The oligos were phosphorylated in a 50 μL reaction containing 1X T4 PNK buffer, 1 mM ATP, 5 mM DTT, 0.038 μM oligos and 10 units T4 PNK. The reaction was incubated at 37 °C for 1 hour, and the enzyme inactivated at 65 °C for 20 minutes.

The PFunkel reaction was prepared in a 0.5 mL eppendorf tube containing 1X PfuTurbo Cx hotstart DNA polymerase buffer, 10 mM DTT, 0.5 mM NAD+, 0.2 mM dNTPs, 1 μL of the kinase reaction, 1 μg (0.75 pmol) of dU-ssDNA template, 2.5 units PfuTurbo Cx hotstart DNA polymerase, and 200 cohesive end units Taq ligase in a total volume of 100 μL. The following denaturation/annealing/extension steps were performed: 95 °C for 2 min, 15 cycles of 95 °C for 30 sec, 55 °C for 45 sec, and 68 °C for 6.5 min. At the 95 °C step of cycles 6 and 11, 1 μL of the kinase reaction was added and mixed in by stirring with the pipette tip. The reaction was then incubated at 45 °C for 15 min for ligation to occur. Then 3.8 pmol of oligo P320 (5:1 molar ratio oligo to template) was added and one more cycle of 95°C for 30 sec, 55°C for 45 sec, and 68°C for 10 min was carried out. The reaction was again incubated at 45 °C for 15 min. Then

10 units of UDG and 30 units of ExoIII were added and incubated at 37 °C for 1 hr followed by an inactivation step at 70 °C for 20 min. The DNA was then purified using the Zymo DNA Clean & Concentrator kit according to the manufacturer's instructions and eluted in 15 µL of DI water. This volume was then vacuum concentrated down to 1-2 µL. For CCM-1 the DNA was electroporated into 50 µL DH5α electrocompetent cells and then incubated with SOC recovery media for 1 hr at 37 °C with shaking at 250 rpm. The entire volume was then plated on a Nalgene Bioassay dish (D4803; 245 mm × 245 mm × 25 mm) containing LB-agar with 50 µg/mL Spec and incubated overnight at 37 °C. For CCM-2, the DNA was used to transform NEB 5-alpha F'I^q competent cells as per the manufacturer's instructions, and then plated on a Nalgene Bioassay dish containing LB-agar with 50 µg/mL Spec, 15 µg/mL tetracycline, and 2 w/v% glucose.

Table 2.1. Reaction conditions for PFunkel using a ssDNA template.

	Site-directed mutagenesis	Multiple-site mutagenesis	Comprehensive codon mutagenesis
Start with	1X pfuTurbo Cx buffer 2.5 units pfuTurbo Cx polymerase 10 mM DTT 0.5 mM NAD^+ 0.2 mM dNTPs 1 µg (0.75 pmol) of dU-ssDNA 1 µl of 3 µM kinased mutagenic oligo 200 cohesive end units Taq ligase 100 µl total volume	1X pfuTurbo Cx buffer 1 µg (0.75 pmol) of dU-ssDNA 2 µl of 6 µM kinased mutagenic oligo mixture 77 µl total volume	1X pfuTurbo Cx buffer 2.5 units pfuTurbo Cx polymerase 10 mM DTT 0.5 mM NAD^+ 0.2 mM dNTPs 1 µg (0.75 pmol) of dU-ssDNA 1 µl of 0.038 µM kinased mutagenic oligo mix 200 cohesive end units Taq ligase 100 µl total volume
Initial step	none	none	95°C for 2 min.
Cycling	1 cycle of 95 °C for 3 min 55 °C for 90 sec 68 °C for 15 min	95°C for 3 min, 55°C for 10 min hold at 55°C. add 10 mM DTT 0.5 mM NAD^+, 0.2 mM dNTPs 2.5 units pfuTurbo Cx (previously heat activated) 200 cohesive end units Taq ligase bringing the total volume to 100 µL 65 °C for 15 min	15 cycles of 95°C for 30 sec 55°C for 45 sec 68°C for 6.5 min. At the 95°C step of cycles 6 and 11, an additional 1 µL of the kinase reaction was added.
Ligation	45°C for 15 min		
add	Add 3.8 pmol of kinased oligo P320 (5:1 molar ratio oligo to template)		
Synthesis of second strand containing mutation	one cycle of 95°C for 30 sec 55°C for 45 sec 68°C for 10 min 45°C for 15 min (ligation)		
Degradation of template and side-products	Shift to 37°C Add 10 units of UDG + 30 units of ExoIII; incubate 1 hour at 37°C; 70 °C for 20 min (heat inactivation)	Shift to 37°C Add 5 units of UDG + 2 units of ExoIII; incubate 30 min at 37°C; 70 °C for 20 min (heat inactivation)	Shift to 37°C Add 10 units of UDG + 30 units of ExoIII; incubate 1 hour at 37°C; 70 °C for 20 min (heat inactivation)
Final step	Purify DNA (optional step to increase number of transformants) and transform		

2.2.6 454 GS FLX high-throughput sequencing

Transformants were recovered from agar plates with LB broth, and plasmid DNA recovered using the Qiagen QIAprep Spin Miniprep kit (27106). The plasmid DNA was

linearized by restriction endonuclease digestion with NdeI. PCR amplicons of each of the three *bla* libraries were created using Titanium Lib-A fusion primers that included a 10-base MID barcode. Each 25 μL PCR reaction had 1-2 ng linearized template DNA, 0.4 μM each primer, 200 μM each dNTP, 1X HF Phusion buffer, and 2 units Phusion high-fidelity polymerase. Cycler conditions were 98 °C for 30 sec, 30 cycles of 98 °C for 30 sec, 55 °C for 30 sec, 72 °C for 30 sec, and then 72 °C for 5 min. PCR products were visualized on an ethidium bromide 1% agarose gel, and then gel purified using the QIAquick Gel Extraction Kit (28706). Amplicons were furthered purified using the Agencourt AMPure XP PCR Purification kit (A63880), to remove short DNA fragments, primers, and primer dimers. DNA concentration was determined using the Quant-iT Picogreen dsDNA Assay kit (P7589). Amplicons from each sub-library were diluted to 10^9 molecules/μL in 1X TE, equal volumes pooled together and then further diluted to 10^7 molecules/μL in DI water. 454 sequencing was performed by Tufts University Core Facility on a Roche 454 GS FLX+ instrument. The sequencing data was then analyzed using the Galaxy open web-based platform (45-47) and custom Matlab scripts.

2.2.7 Identification of adaptive codon substitutions for tazobactam resistance in *TEM-1*.

Library CCM-2 was plated at a density of about 500 CFU/cm^2 (non-selective conditions) on LB-agar plates supplemented with 50 μg/ml Spec, 300 μM IPTG, 100 μg/ml ampicillin and 4 or 6 μg/ml tazobactam. Plates were incubated at 37 °C for 17 hours. The tazobactam concentration chosen was 1.3 or 2-fold higher than the concentration at which cells bearing wildtype *TEM-1* could grow effectively. Large colonies on the plates were chosen at random for sequencing.

Selected single base mutations were re-introduced into TEM-1 by site-directed

PFunkel mutagenesis on the 20 µl volume scale. The MIC for ampicillin and piperacillin of the mutants was assessed with and without 6 µg/ml tazobactam by spotting 10^4 CFU on Mueller-Hinton agar plates containing 50 µg/ml Spec, 300 µM IPTG, and √2-fold increments of either ampicillin or piperacillin. Plates were incubated at 37 °C for 12 hrs.

2.2.8 Preparation of uracil-containing dsDNA template for phage-less PFunkel

A 10 mL LB culture of CJ236 cells with the pSkunk3-bla plasmid was incubated overnight at 37 °C with shaking at 250 rpm. Plasmid dU-dsDNA was then isolated using the Qiagen QIAprep Spin Miniprep kit (27106) and the concentration quantified using a Nanodrop ND-1000 spectrophotometer.

2.2.9 Site-directed PFunkel mutagenesis using a plasmid dsDNA template

All steps were performed in a pre-programmed Eppendorf Mastercycler personal thermocycler. A mutagenic oligo (5'-gacaccacgatgcatgcagcaatggc) encoding a c542a mutation in *TEM-1* was phosphorylated in a 50 µL reaction containing 1X T4 PNK buffer, 1 mM ATP, 5 mM DTT, 1.5 µM oligo and 10 units T4 PNK. The reaction was incubated at 37 °C for 1 hour and the enzyme inactivated at 65°C for 20 minutes.

The PFunkel reaction was prepared in a 0.5 mL tube containing 1X PfuTurbo Cx hotstart DNA polymerase buffer, 10 mM DTT, 0.5 mM NAD+, 0.2 mM dNTPs, 1 µL of the kinase reaction, 1 µg (0.38 pmol) of dU-dsDNA template, 2.5 units PfuTurbo Cx hotstart DNA polymerase, and 200 cohesive end units Taq ligase in a total volume of 100 µL. The free Mg^{2+} concentration should be maintained between 0.5-2.5 mM since low concentration reduces polymerase fidelity while high concentration leads to nonspecific annealing of oligos (44). The volume of kinase reaction added should therefore be minimized to maintain Mg^{2+} concentration in the mutagenesis reaction close to the 2 mM

Mg^{2+} provided in the 1X polymerase buffer. The following denaturation/annealing/extension/ligation steps were performed: 95 °C for 3 min, 55 °C for 90 sec, 68 °C for 15 min and 45 °C for 15 min. Next, 10 units of UDG and 30 units of Exo III were added and the reaction was incubated at 37 °C for 1 hr followed by an inactivation step at 70 °C for 20 min. A total of 3.8 pmol of oligo P320 (5'-P-gcagaaattcgaaagcaaattcgac) was added and one more cycle of 95 °C for 30 sec, 55 °C for 45 sec, 68 °C for 10 min and 45 °C for 15 min was performed. The DNA was purified using the Zymo DNA Clean & Concentrator kit according to the manufacturer's instructions and eluted in 15 µL of DI water. This solution was vacuum concentrated down to 1-2 µL, electroporated into 50 µL DH5α electrocompetent cells, which were incubated with SOC recovery media for 1 hr at 37 °C with shaking at 250 rpm. The transformation was plated on LB-agar with 50 µg/mL Spec and incubated overnight at 37 °C.

2.2.10 Multi-site PFunkel mutagenesis using a plasmid dsDNA template

Four oligos were designed to introduce the four mutations A42G, E104K, M182T, and G238S in the *bla* gene with respective sequences: 5'-gatcagttgggtgga cgagtgggttac, 5'-ctcagaatgacttggttaagtactcaccagtcacag, 5'-gtgacaccacgacgcctgcagcaatggcaacaac, 5'-gctgataaatctggagccagtgagcgtgggtctcg. These oligos were combined in equimolar amounts and phosphorylated in a 50 µL reaction containing 1X T4 PNK buffer, 1 mM ATP, 5 mM DTT, 3 µM total oligo and 10 units T4 PNK. The reaction was incubated at 37 °C for 1 hour and the enzyme inactivated at 65 °C for 20 minutes.

The annealing reaction was prepared in a 0.5 mL eppendorf tube containing 1X PfuTurbo Cx hotstart DNA polymerase buffer, 2 µL of kinase reaction, and 1 µg of dU-

dsDNA template in a total volume of 77 µL. The annealing was performed by heating to 95 °C for 3 min, then 55 °C for 10 min, and holding at 55 °C.

Meanwhile, in a separate PCR tube, 1X PfuTurbo Cx hotstart DNA polymerase buffer and 2.75 units of PfuTurbo Cx hotstart DNA polymerase were combined in a total volume of 5.5 µL. The hotstart polymerase was heat activated by heating to 95 °C for 3 min.

After the annealing step, 10 mM DTT, 0.5 mM NAD+, 0.2 mM dNTPs, 5 µL of the activated polymerase solution, and 200 cohesive end units Taq ligase was added bringing the total volume to 100 µL. The reaction was mixed by slowly and gently pipetting up and down. Extension and ligation of the mutant strand was performed at 65 °C for 15 min and 45 °C for 15 min. Five units of UDG and 2 units of ExoIII were added and the mixture was incubated at 37 °C for 1 hr followed by an inactivation step at 70 °C for 20 min. A total of 3.8 pmol of oligo P320 was added and one more cycle of 95 °C for 30 sec, 55 °C for 45 sec, 68 °C for 10 min and 45 °C for 15 min was performed. The DNA was purified using the Zymo DNA Clean & Concentrator kit according to the manufacturer's instructions and eluted in 15 µL of DI water. This solution was vacuum concentrated down to 1-2 µL, electroporated into 50 µL DH5α electrocompetent cells and then incubated with SOC recovery media for 1 hr at 37 °C with shaking at 250 rpm. The transformation was plated on LB-agar with 50 µg/mL Spec and incubated overnight at 37 °C.

Table 2.2. Reaction conditions for PFunkel using a dsDNA template

	Site-directed mutagenesis	Multiple-site mutagenesis
Start with	1X pfuTurbo Cx buffer 2.5 units pfuTurbo Cx polymerase 10 mM DTT 0.5 mM NAD^+ 0.2 mM dNTPs 1 µg (0.38 pmol) of dU-dsDNA 1 µl of 1.5 µM kinased mutagenic oligo 200 cohesive end units Taq ligase 100 µl total volume	1X pfuTurbo Cx buffer 1 µg (0.38 pmol) of dU-dsDNA 2 µl of 3 µM kinased mutagenic oligo mixture 77 µl total volume
Initial step	none	none
Cycling	1 cycle of 95 ^oC for 3 min 55 ^oC for 90 sec 68 ^oC for 15 min	95°C for 3 min, 55°C for 10 min hold at 55°C. add 10 mM DTT 0.5 mM NAD^+, 0.2 mM dNTPs 2.5 units pfuTurbo Cx (previously heat activated) 200 cohesive end units Taq ligase bringing the total volume to 100 µL 65 ^oC for 15 min
Ligation	45°C for 15 min	
Degradation of template and side-products	Shift to 37°C Add 10 units of UDG + 30 units of ExoIII; incubate 1 hour at 37°C; 70 ^oC for 20 min (heat inactivation)	Shift to 37°C Add 5 units of UDG + 2 units of ExoIII; incubate 1 hr at 37°C; 70 ^oC for 20 min (heat inactivation)
add	Add 3.8 pmol of kinased oligo P320 (10:1 molar ratio oligo to template)	
Synthesis of second strand containing mutation	one cycle of 95°C for 30 sec 55°C for 45 sec 68°C for 10 min 45°C for 15 min (ligation)	
Final step	Purify DNA (optional step to increase number of transformants) and transform	

2.3 RESULTS

2.3.1 Limitations of Kunkel mutagenesis

We sought to minimize the occurrence of wild-type sequences in Kunkel mutagenesis, which has been reported to be as high as 30-50% (28). We postulated that wild-type sequences arise for several reasons. First, the low operating temperature of the second strand synthesis step allows any contaminating short DNA fragments in the single

stranded DNA prep, termed "junk" DNA, to prime the single stranded DNA. Such synthesis can either create a wild-type double stranded product or poison a mutation-bearing product by creating reaction side-products that possess a nick or a displaced strand (48). Additionally, at lower temperatures, the mutagenic oligos are more prone to anneal non-specifically to the template. Such reaction side-products that are not in the cccDNA form are prone to degradation by cellular nucleases, removing the mutation. The presence of junk DNA is apparent from DNA gels of reaction products in which no mutagenic oligonucleotides were added, yet higher molecular weight products are produced (49). Another postulated reason for the high occurrence of wild-type sequences in Kunkel mutagenesis is the repair of the mutation by mismatch-repair machinery or repair of the uracil-containing template strand in the cell after transformation.

2.3.2 Overview of PFunkel mutagenesis using a ssDNA template

The in vitro reaction steps of PFunkel mutagenesis (Figure 2.1, Table 2.1) are designed to eliminate products other than the desired mutated cccDNA plasmid molecules, resulting in high mutational efficiencies. Other than the initial kinase reaction to phosphorylate the mutagenic oligonucleotides, all reaction steps are conveniently performed in the same tube with no DNA purification required except as an optional final step to improve transformation efficiency. PFunkel is conveniently performed in a thermocycler.

Uracil-containing ssDNA is produced by propagating phagemid DNA containing the DNA to be mutated in an *E. coli dut-1 ung-1* host, then infecting the culture with M13 helper phage and harvesting ssDNA from the resulting phage particles. *E. coli dut-1 ung-1* strains express a heat-sensitive dUTPase that has 5% of wildtype activity at 25 °C but

<1% at 37 °C, and are deficient in uracil DNA glycosylase activity (50). The result of these mutations is the accumulation of high levels of intracellular dUTP that becomes incorporated in DNA in place of dTTP during DNA synthesis and is not removed due to the lack of UDG activity. ssDNA preparation takes only a day and requires no special laboratory equipment or highly specialized training (49).

To minimize second strand synthesis originating from junk DNA annealing to the uracil-containing ssDNA template, we shifted the operating temperature from the 25-37 °C to 55-95 °C. This operating temperature required a high-fidelity thermostable polymerase capable of using a uracil-containing template. Additionally, a polymerase lacking strand displacement activity would be advantageous for creating multiple mutations simultaneously at different sites in a gene. The only commercially available polymerase that met these criteria was PfuTurbo Cx hotstart DNA polymerase (Agilent), a variant of *Pfu* polymerase with the V93Q mutation (51). This mutation inactivates the uracil-binding pocket of the enzyme that would normally cause it to stall at uracil bases. At ≤68 °C PfuTurbo Cx hotstart DNA polymerase does not strand-displace but still maintains significant polymerase activity (52). Taq ligase was chosen due to its effectiveness in ligating DNA nicks, robust activity from 45°C to 65 °C, and ability to withstand many rounds of temperature cycling.

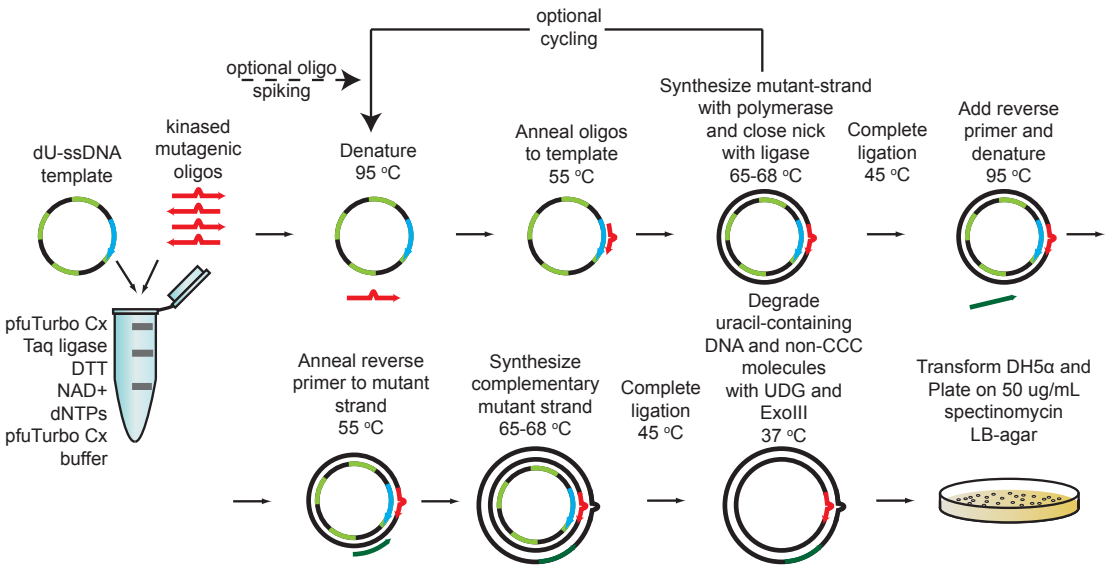


Figure 2.1. Schematic of PFunkel mutagenesis using a ssDNA template. The basic protocol is depicted. For multiple-site mutagenesis, the addition of the polymerase, dNTPs, ligase, DTT, and NAD+ is delayed until after the first annealing step. For comprehensive codon mutagenesis, the ratio of oligo to template is kept low to minimize multiple mutations in a single reaction product. Cycling with occasional spiking of additional mutagenic oligos improves the reaction yield.

In order to create the dsDNA product with the designed mutation on both strands, an excess of a 'reverse' oligonucleotide that anneals outside of the gene on the newly created mutant strand is added, such that it primes synthesis of a new complementary strand that encodes the desired mutations and displaces the uracil-containing template. Treatment with UDG acts to excise the uracil bases from the original template strand leaving apyrimidinic (AP) sites. Treatment with exonuclease III (ExoIII), which has both AP-site endonuclease and 3'->5' exonuclease activity (53), acts to create nicks at the AP sites and then digests the template strand at the nicks and from any 3' end in the context of dsDNA.

2.3.3 Site-directed PFunkel mutagenesis

A mutagenic oligonucleotide encoding a c542a (P183H in the protein) mutation in

the TEM-1 ß-lactamase (*TEM-1*) gene was first 5' phosphorylated in a kinase reaction. The phosphorylated oligo was then combined with the ssDNA uracil-containing template in molar ratio of 4:1 together with the polymerase and ligase. The incubation temperatures were cycled to perform a denaturing, annealing, extension, and ligation step to complete the mutated second strand and seal the nick. A second primer that annealed to the new strand outside the gene was added to the reaction, and the denaturing, annealing, extension, and ligation steps were repeated. Exo III and UDG were then added to the reaction to remove the template and undesired side-products. All steps for this procedure took about 3 hrs to complete.

A transformation of 5 μL of the unpurified reaction with 100 μL of chemically competent cells yielded over 1000 transformants, illustrating that DNA purification is not necessary. The remaining DNA was purified using a spin column and $1/15^{th}$ of the product was electroporated into electrocompetent DH5α *E. coli* yielding 533,000 transformants. Sequencing of the TEM-1 gene from 23 colonies showed that all 23 (100%) contained the c542a mutation encoded by the oligo. No undesired mutations were observed. We further substantiated the high mutational efficiency of our method using eleven different oligos encoding either a 1 or 2 base substitution at different locations of the gene (Table 2.3).

Table 2.3. Results of additional testing of PFunkel site-directed mutagenesis and multi-site mutagenesis using a single-stranded DNA template.

Type of PFunkel	Oligonucleotides added in reaction[a] Mutation(s) Intended	Number of correct sequences[b]
Site-directed Mutagenesis	M69L	2 of 2
	Y105S	2 of 2
	Y105D	2 of 2[c]
	Y105N	2 of 2
	S235T	2 of 2
	R244S	2 of 2
	N276D	2 of 2
	A42G	2 of 2
	E104K	2 of 2[d]
	M182Q	2 of 2
	G238A	2 of 2
Multi-site Mutagenesis	A42G, E104K	2 of 2
	A42G, M182Q	2 of 2
	A42G, G238A	2 of 2
	E104K, M182Q	2 of 2
	E104K, G238A	1 of 2
	M182Q, G238A	1 of 2
	A42G, E104K, M182Q	1 of 2
	A42G, E104K, G238A	1 of 2
	A42G, M182Q, G238A	1 of 2
	E104K, M182Q, G238A	1 of 2
	A42G, E104K, M182Q, G238A	2 of 2

[a] M69L encodes atg69ctg (5'- cccgaagaacgttttccaatgctgagcacttttaaa-3')
Y105S encodes tac105tcc (5'- tgacttggttgagtcctcaccagtcacaga-3')
Y105D encodes tac105gac (5'- tgacttggttgaggactcaccagtcacaga-3')
Y105N encodes tac105aac (5'- tgacttggttgagaactcaccagtcacaga-3')
S235T encodes tct235act (5'- attgctgataaaactggagccggtgagc-3')
R244S encodes cgc244agc (5'- gagcgtgggtctagcggtatcattgca-3')
N276D encodes aat276gat (5'- atggatgaacgagatagacagatcgctgaga-3')
A42G encodes gca42ggg (5'- gatcagttgggtgggcgagtgggttac-3')
E104K encodes gag104aag (5'- ctcagaatgacttggttaagtactcaccagtcacag -3')
M182Q encodes atg182cag (5'- cgtgacaccacgcagcctgcagcaatg -3')
G238A encodes ggt238gca (5'- aaatctggagccgcagagcgtgggtct -3')
[b] The number of clones with all intended mutations out of the total number of clones sequenced.
[c] One clone had an additional, unintended point mutation.
[d] One clone had an additional, unintended mutation found near the desired mutation, presumably resulting from the synthesized mutagenic oligo possessing a misincorporated base.

2.3.4 Multi-site PFunkel mutagenesis

Existing methods for site-directed mutagenesis at multiple distal sites simultaneously either have complex and multi-step procedures or have not been

demonstrated to be efficient enough for library construction (32,54). PFunkel was designed in part to allow efficient construction of libraries in which site-saturation mutagenesis (or any user-defined mutational composition) can be performed at multiple sites simultaneously in a single reaction.

For simultaneous introduction of mutations at multiple distant sites in a gene, the basic PFunkel protocol is modified to increase the frequency of multiple mutations. The polymerase is added after the annealing of the mutagenic oligos. The rationale for the delayed addition of polymerase is to prevent a bias for mutations that result from oligos that anneal efficiently. DNA synthesis from such early annealing oligos might proceed to regions of the gene where other oligos are intended to anneal before the oligos for those locations have a chance to anneal, thus decreasing the frequency of multiple mutations in the resulting transformants. We also reduce the extension temperature to 65°C, to better ensure that PfuTurbo Cx hotstart DNA polymerase does not strand displace. Strand displacement of a strand created from one mutagenic oligo by a strand being synthesized starting from a second mutagenic oligo would reduce the frequency of multiple mutations. We do not know whether the shift of the extension temperature from 68°C to 65°C provides any benefit as we have not performed any direct comparison.

To demonstrate multiple-site mutagenesis using PFunkel, we synthesized four mutagenic oligos designed to create site-saturation libraries of four codons in different regions of the *TEM-1* gene simultaneously. The oligos encoding NNN at codon positions A42, E104, M182, and G238 were combined with the ssDNA template such that each oligo was present in an oligo to template molar ratio of 4:1. Electroporation of the entire reaction product after spin column purification yielded 5.8 million transformants.

Sequencing of the *TEM-1* gene of 10 colonies showed that 7 variants had mutations at all 4 designated codon positions, 2 had mutations at 3 positions, and 1 had mutations at 2 positions (Table 2.4). Twenty-nine of the 35 codon substitutions were unique and no undesired mutations were observed. We further substantiated multi-site PFunkel by constructing 11 different double, triple, or quadruple mutants at 73% efficiency (Table 2.3). The error rate for single or multi-site PFunkel mutagenesis was ~5 x 10^{-5}, higher than expected based on the error rate of PfuTurbo Cx in a PCR reaction (see section 2.3.8).

Table 2.4. Mutations in 10 clones of the naïve multi-site library

	Ambler Position[a]							
	42		104		182		238	
Colony	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
TEM-1	gca	A	gag	E	atg	M	ggt	G
1	ggc	G	-	-	cag	Q	tgg	W
2	agc	S	ggg	G	gga	G	-	-
3	cgg	R	ggg	G	cgg	R	aga	R
4	tgc	C	ggg	G	agg	R	cgg	R
5	tgc	C	gcg	A	ggg	G	gtg	V
6	ggc	G	tgg	W	ccg	P	ctg	L
7	ggc	G	ggg	G	ttg	L	gac	D
8	ggt	G	ggg	G	gca	A	atc	I
9	gcg	A	ggt	G	cat	H	atc	I
10	-	-	-	-	-	-	tcc	S

[a] oligo for position 42: 5'-gatcagttgggtnnncgagtgggttac-3'
oligo for position 104: 5'-gaatgacttggttnnntactcaccagtcac-3'
oligo for position 182: 5'-cgtgacaccacgnnncctgcagcaatg-3'
oligo for position 238: 5'-aaatctggagccnnngagcgtgggtct-3'

2.3.5 Comprehensive codon mutagenesis

We next used PFunkel to create a library designed to encompass all possible single codon substitutions in the *TEM-1* gene (287 codons x 63 possible codon substitutions at each codon = 18,081 desired mutants). We did not desire library members with more than one codon substituted. Such a library is the equivalent of performing site-

saturation mutagenesis at all positions in the gene simultaneously. The advantage of PFunkel is that one does not have to perform 287 separate mutagenesis reactions or 287 separate gene syntheses to create this library. The library would also be much closer to a true random mutagenesis library than one created by error prone PCR, which is biased towards certain base substitutions made by the polymerase and certain amino acid substitutions accomplished by single base mutations.

The 287 degenerate mutagenic oligos (one for each of the 287 codons to be mutated) were designed *in silico* using a Matlab script (Figure 2.2). The oligos were purchased in desalted 96-well format using machine-mixed degenerate bases and pooled. To minimize the occurrence of multiple mutations, the total oligo to ssDNA template ratio was kept low (1:20), which makes two oligos annealing to the same ssDNA template unlikely. To increase the yield and efficiency of the reaction, we implemented a cycling reaction of denaturing, annealing, and extension to allow multiple chances for each oligo to productively anneal. Fifteen cycles were performed with additional oligos spiked in at the 6^{th} and 11^{th} cycle. Additional cycles and oligo additions can be performed if larger libraries are desired. Our scheme is analogous to that of a discontinuous fed batch reactor – a reaction strategy to minimize undesirable side products that occur with a high concentration of one of the reactants (55).

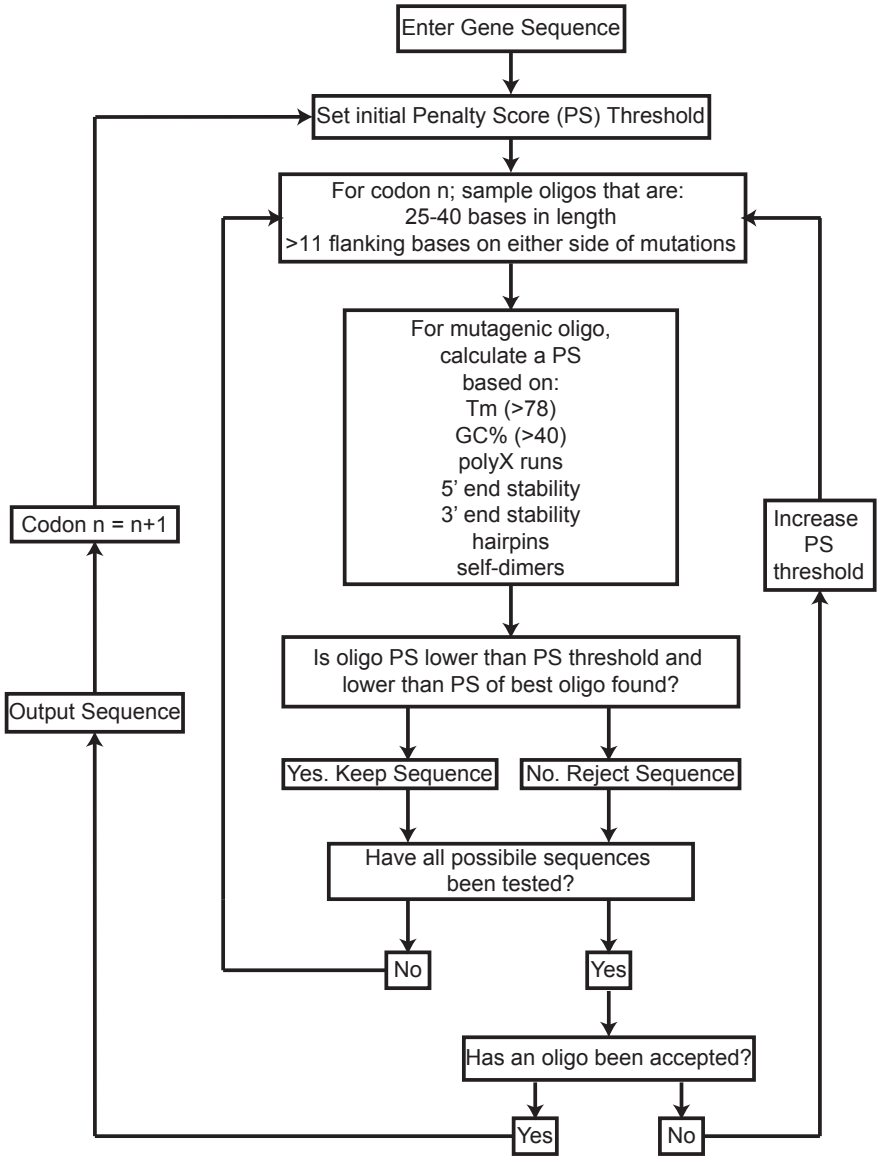


Figure 2.2. Schematic of the Matlab algorithm for designing the mutagenic oligos for comprehensive codon mutagenesis. For each gene position to be randomized, the algorithm scans through many possible oligos, assigns each a score based on specific guidelines, and then selects the best scoring oligo sequence. Published design criteria [7] along with early experimental data were used to develop the following oligo criteria: a) the oligo length can vary from 27 to 40 bases; b) the mismatched bases must be flanked by ≥12 bases on each side; c) the T_m must be ≥ 62 °C; d) the GC content must be ≥ 40%; e) oligos with a stable 5' end and an unstable 3' end are favored to prevent non-specific annealing and extension; and f) oligos with polynucleotide repeats, hairpin structures, and a propensity for dimerization are penalized. Each oligo is designed to replace a different codon in the *bla* gene with a random sequence (NNN). The script can be easily modified for designing other types of libraries.

Although in principle the library could be created in a single tube, we divided the library into thirds corresponding to each 1/3 of the gene in order to facilitate characterization of the library by 454-GS-FLX Titanium sequencing, which has a read length of ~400 bp for the sequencing of amplicons from pools of DNA. Transformation of the entire reaction product yielded ~500,000 transformants for each library. Sequencing of 30 members of each library indicated that the libraries mostly consisted of single codon substitutions (87%) with the remainder being wildtype (13%) (Table 2.5). No clones with multiple mutations were observed. Two of the sequences contained a single mutation outside the region subjected to mutagenesis, which we attribute to polymerase error. We postulate that additional rounds of cycling and mutagenic oligo addition would lower the fraction of wild-type sequences closer to the theoretical minimum of 1.6% (i.e. 1/64 of the NNN containing oligos encode the wildtype codon). These three libraries collectively were named CCM-1.

2.3.6 454 GS FLX high-throughput sequencing analysis of the comprehensive codon substitution library

Barcoded amplicons from the three CCM-1 libraries were created by PCR and pooled. Additionally, barcoded amplicons created from the wildtype *TEM-1* gene were added to the pool as a control for sequencing errors. We obtained 787,488 reads that passed quality filtering, with a median length of 354 bases. A total of 99% of the reads spanned the entire mutated region of the amplicon. The reads of the library DNA displayed a higher frequency of both wildtype (26%) and multiple mutations (17%) at the expense of single codon mutations (57%) as compared to the Sanger sequencing of 90 clones. However, this was determined to be an artifact of the amplicon preparation known as “PCR jumping,” a well-documented occurrence during PCR amplification of

highly-identical, heterogeneous template sequences in which chimera PCR products are produced (56,57). We confirmed this was the case by Sanger sequencing of 28 individual PCR amplicons of which 36% had no mutations, 50% had one codon mutation, and 14% had multiple codon mutations (Table 2.5). This closely matched the proportions in the 454 sequencing. The sequencing of wild-type *TEM-1* indicated that the sequencing error rate (0.035 codon substitutions per read) was much less than the frequency of codon substitutions observed in the reads of the library DNA (0.94 per read). We conclude that 96% of the codon substitutions observed in the 454 sequencing reads are present in the library, with the remainder being sequencing errors. Of the codon substitutions present in the library, ≤3% are present in library members with multiple mutations (based on Sanger sequencing).

In the 454 sequencing reads of the library we observed 97.0% of the 18,081 intended codon substitutions at least once. In the worst-case scenario in which sequencing errors are assigned to mutations with the lowest numbers of occurrences, 84.8% of the 18,081 possible mutants are present in the library. If sequencing errors are evenly distributed across all codon substitutions, 96.4% are present. In the best-case scenario in which errors are assigned to mutants that are highly represented, 96.8% of the 18,081 mutants are present. We believe that the true coverage of the library lies between 96.4% and 97% and likely closer to 97%, since 454 sequencing is known to exhibit sequence-dependent common errors. Among the 72 sequencing errors in the reads of wild-type *TEM-1*, one particular substitution appeared five times and five codon substitutions appeared twice. More extensive sequencing of wild-type *TEM-1* would be necessary to accurately determine the frequency at which each of the 18,081 possible

codon substitutions appear because of sequencing error, and thus the true frequency of each codon substitution in the library.

Table 2.5. Statistics of comprehensive codon mutagenesis library CCM-1.

	Expected in an ideal library	Sequencing of individual clones of the library	Sequencing of PCR amplicons used in 454 sequencing	454 sequencing of the library	454 sequencing of *TEM-1*
Sequences		90 clones	28 clones	787,488 reads	2040 reads
Percent of reads that cover entire gene segment		100%	100%	98.95%	99.75%
Number of mutated codons in all sequences		78 + 2[a]	22	738,615	72
Mean mutated codons per sequence	0.9844	0.87	0.79	0.94	0.035
Percent of clones/reads with					
No mutations	1.56%	13.33%	35.71%	26.17%[b]	96.9%
One mutation	98.44%	86.67%	50.00	56.71%	2.75%
Multiple mutations	0.00%	0.00%	14.29%	17.12%[b]	0.034%
Percent of mutated codons with					
1 base substitution	14.29%	22.50%	22.72%	31.97%	86.11%
2 base substitution	42.86%	47.50%	59.10%	41.84%	13.89%
3 base substitution	42.86%	30.00%	18.18%	26.20%	0.00%
Percent of possible codon substitutions observed					
1 base substitution				99.96%	
2 base substitutions				97.70%	
3 base substitutions				95.33%	
All substitutions				97.01%	

[a]two mutations were identified outside the region targeted for mutagenesis

[b]most of the reads with multiple mutations and about 50% of the reads with no mutations result from PCR jumping during amplicon creation. The library mostly is comprised of members with one mutation as indicated in sequencing of individual clones.

Our analysis indicates that 96-97% of the 18,081 desired codon substitutions are present in the library, and ≥97% of library members with a codon substitution contain only one codon substitution (Table 2.5). The frequency of codon substitutions observed as a function of gene position shows that a few positions were hotspots for substitutions and that the frequency of codon substitutions has a broad distribution (Figure 2.3). Presumably the occurrence of hotspots reflects the suitably of the respective oligos for this mutagenesis technique. Codon substitutions with a single bp change were observed at about twice the expected frequency, and this comes at the expense of fewer codon substitutions with three bp changes (Table 2.5). A portion of the bias towards single base substitutions is likely due to polymerase errors during library construction, polymerase errors during the PCR-based amplicon preparation for sequencing, and 454 sequencing errors, all of which would be primarily single base substitutions. The remainder of the bias may reflect the increased mismatch between the mutagenic oligo and the template for codon substitutions with three mutated bases. Still, although codon substitutions with three bp changes may be somewhat disfavored, we observed 95% of the 7749 designed 3-bp change codon substitutions in the 454 sequencing results (Table 2.5). Since 454 sequencing errors with three bp changes in a codon are likely very rare, we believe most if not all of these 3-base substitutions are present in the library.

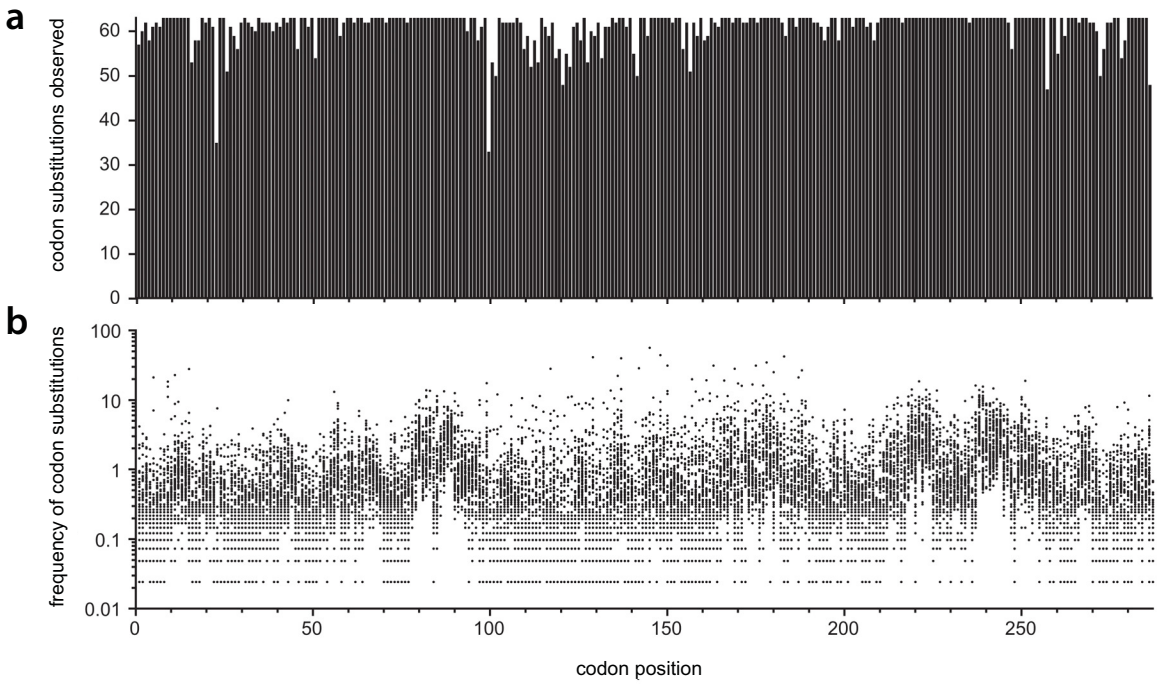


Figure 2.3. Completeness and frequency of codon substitutions observed in 454 sequencing of the comprehensive codon mutagenesis library of *TEM-1*. (**a**) Number of the 63 possible codon substitutions observed and (**b**) frequency of codon substitutions observed as a function of position in the gene. For each of the 287 codons of *TEM-1* the frequency of each of the 63 possible codon substitutions is shown, except for the 3% of the 18,081 codon substitutions that were not observed. The frequency is based on 454 sequencing in which 738,615 codon substitutions were observed in 787,488 reads. The frequency is normalized to the frequency that would occur if all substitutions were evenly distributed among the 18,081 possible substitutions (i.e. frequency = 1.0 means that the substitution was observed 738,615/18,081 = 41 times). The number of codon substitutions observed resulting from sequencing errors is small (~4% of the 738,615 codon substitutions observed).

Both the Sanger and the 454 sequencing indicate that G's are present in mutated codons 2.3 times more frequently than any one of the other three bases (Table 2.6). The high frequency of G's is also apparent in the sequences of naïve members of the multi-site mutagenesis library (Table 2.4), which used four specific primers from the set of 287. The distribution of the frequency of the substituted codons strongly reflects this bias (Figure 2.4A) whereas the distribution of the frequency of codons substituted into does

not (Figure 2.4B). Since *TEM-1* has roughly an equal frequency of all bases, we conclude that this bias results from a 2.3-fold bias for incorporation of G's during the synthesis of the machine-mixed degenerate oligonucleotides. This bias contributed to the underrepresentation of certain mutations, as the frequency of G's in codon substitutions not observed in the 454 sequencing was 0.068.

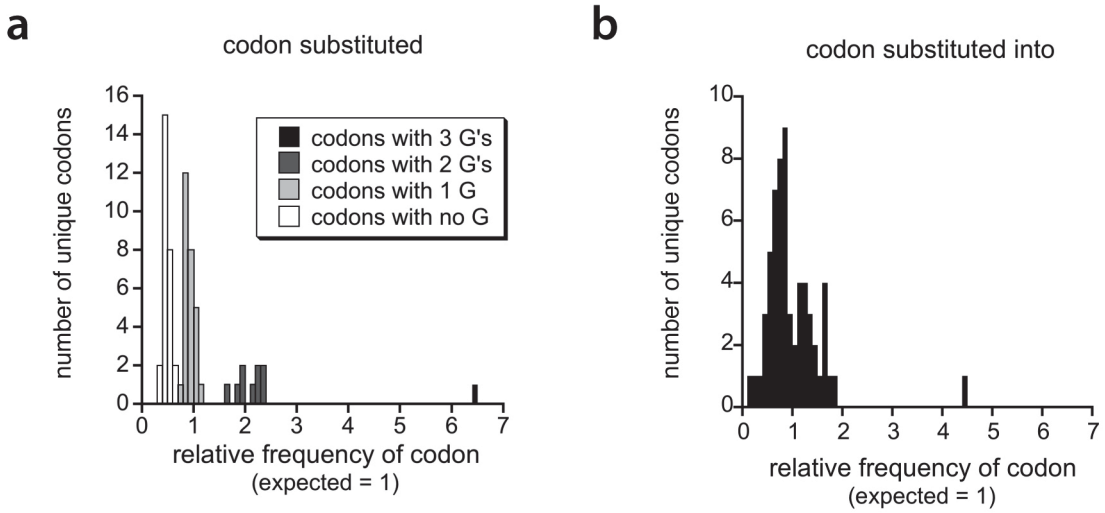


Figure 2.4. Distribution of codon frequencies. Distribution of the frequency of the type of (**a**) codon substitutions and (**b**) codons substituted into the comprehensive codon mutagenesis library CCM-1. The frequency is normalized to that expected if all codon substitutions occurred with equal frequency. The codon substitutions are color coded as to the number of G's in the substituted codon. *TEM-1* lacks three codons (TAG, TGA, AGG) so those codons are not included in the codons substituted into.

Table 2.6. Percent of bases in mutated codons in the comprehensive codon mutagenesis library CCM-1.

Base	Expected in an ideal library	Sequencing of 90 individual clones of the library	454 sequencing of library	454 sequencing of *TEM-1* (i.e. sequencing errors)
G	25.00%	46.25%	43.26%	26.85%
A	25.00%	17.08%	18.84%	18.52%
T	25.00%	18.33%	18.81%	24.54%
C	25.00%	18.33%	19.09%	30.09%

2.3.7 Construction and characterization of comprehensive codon mutagenesis library CCM-2.

To confirm that the bias for G's resulted from their overrepresentation in the mutagenic oligos, we constructed a second set of three libraries (CCM-2) using a second set of degenerate mutagenic oligos that were synthesized using a hand-mixed ratio of bases (instead of machine mixed). The three libraries were transformed into NEB 5-alpha F'I^q cells, which contain the $lacI^q$ repressor to better repress expression to avoid any bias when propagating the library. Sequencing of 30 members of each library revealed 8.9% wildtype, 83.1% single codon substitutions in the targeted region, 1.1% with a single mutation outside the targeted region, and 6.7% multiple mutations (3 of 6 had two-mutations in the targeted region; 3 of 6 had one codon substitution in the targeted region and the second mutation in a non-targeted region). The frequency of bases substituted in the designed mutations of CCM-2 (27.5%:26.6%:23.0%:23.0% for G:A:C:T) was much more even than in CCM-1. The ratio of 1-base:2-base:3-base substitutions in the targeted region was 27.7%:42.2%:30.1%.

2.3.8 PFunkel error rate

PfuTurbo Cx hotstart DNA polymerase has a published error rate of 1.3×10^{-6} in a PCR reaction using a double-stranded template (44). For site-directed and multi-site mutagenesis using a single-stranded template we observed three unintended mutations outside the region of the mutagenic oligo in 77 sequencing reactions of the 861 bp *TEM-1* gene, which corresponds to an error rate of 4.5×10^{-5}. For the comprehensive codon mutagenesis, we observed 6 mutations outside the target region in CCM-1 and CCM-2, which corresponds to an error rate of 5.8×10^{-5}. These error rates are 35- and 45-fold higher than PfuTurbo Cx hotstart DNA polymerase's error rate. We speculate that the

elevated error rate results from deviations from the recommended PfuTurbo Cx reaction buffer and/or degradation of the ssDNA template at 95°C. All the observed unintended mutations can be explained by cytosine deamination (~2 x 10^{-7} events/sec at 95°C in ssDNA (58)) leading to G:C->A:T transitions or depurination (~4 x 10^{-7} events/sec at 95°C in ssDNA (59,60)) which can lead to various mutations. When ssDNA is used as the template, we speculate that the 95°C incubation step is not essential for PFunkel and that elimination of this step would lower the error rate.

2.3.9 Identification of adaptive codon substitutions in *TEM-1* that confer increased tazobactam resistance with a single amino acid substitution.

An extensive knowledge of the possible molecular determinants of bacterial resistance to ß-lactam antibiotics and ß-lactamase inhibitors would inform the development and implementation of new antibiotics and inhibitors. We identified adaptive codon substitutions in *TEM-1* conferring increased resistance to the ß-lactamase inhibitor tazobactam, which is used clinically in combination with the extended spectrum ß-lactam antibiotic piperacillin in the drug Tazocin/Zosyn. We identified these adaptive mutations from library CCM-2 – a second comprehensive codon mutagenesis library we constructed that lacked the oligonucleotide-derived bias for G's in the substituted codon observed in CCM-1 (see section 2.3.6 and Figure 2.4). We subjected CCM-2 to a selection for an increase in resistance to tazobactam. Under the selective conditions, wildtype survived at a frequency of about 10^{-3}. Sequencing of 279 colonies revealed 120 unique non-wildtype sequences. Since any particular amino acid substitution is relatively rare in the library, we used the criteria that an amino acid substitution had to be observed twice for us to categorize it as potentially adaptive in nature.

Table 2.7. Potential adaptive amino acid substitutions in *TEM-1* identified from genetic selections for tazobactam resistance codon substitutions.

Ambler position[a]	Amino acid substitutions		Occurrences[d]	Codon coverage[e]
	Clinically observed[b]	This study[c]		
I13	–	**L**	2	2 of 6
L21	F, I	Q	2	1 of 2
M69	L, I, V	**L**	30	6 of 6
Q90	–	A	2	1 of 4
Y105	–	G	14	4 of 4
		S*	10	3 of 6
		A	7	2 of 4
		D	5	2 of 2
		N	5	2 of 2
		W	4	1 of 1
		T	2	2 of 4
R120	G	E	3	1 of 2
S124	N	Q	2	1 of 2
T128	–	E	2	1 of 2
T140	–	G	2	1 of 4
E147	–	**G**	2	2 of 4
W165	R, C, G	Y	4	2 of 2
S235	–	**T**	8	3 of 4
T265	M	**M**	4	1 of 1

[a] Numbering according to Ambler et al. (61).
[b] Amino acid substitutions observed in natural alleles of *TEM-1* with increased resistance to ß-lactam antibiotics or ß-lactamase inhibitors (http://www.lahey.org/studies/temtable.asp). Amino acid substitutions underlined are found in alleles with increased inhibitor resistance (62).
[c] Amino acid substitutions in bold were observed with a single base change in the codon. * means that although the amino acid substitution can occur with a single base change, such a change was not observed here.
[d] Of the amino acid substitution in this study.
[e] For the amino acid substitutions found in this study, the number of unique codons observed out of the possible number of unique codons is reported.

The set of these potentially adaptive substitutions (Table 2.7) overlapped one (M69L) but not other mutations previously found in alleles that increase tazobactam resistance, most notably R244S and N276D (63). In addition, we identified 18 new, potentially adaptive amino acid substitutions, the most prevalent of which were 8

different amino acid substitutions at Y105 and the S235T mutation. The Y105 S/D/N and S235T mutations can occur with a single base change and are the most likely to appear naturally. We introduced these four mutations by single base substitution into *TEM-1* and compared the ampicillin, piperacillin, and tazobactam resistance of these alleles to previously known tazobactam resistance alleles (Figure 2.5). We find that all four provide higher resistance to ampicillin in the presence of tazobactam than R244S and N276D, suggesting that our selection was too strong to identify R244S and N276D. The Y105N, Y105S, and S235T alleles show significant inhibitor resistance for both ampicillin and piperacillin hydrolysis – at or above that of the M69L allele, which is the most resistant allele observed to date for the piperacillin/tazobactam combination (63). We predict that Y105N, Y105S, and S235T have the potential to emerge in the clinic. Their non-emergence to date, and the fact that they were not identified in previous selections for tazobactam resistance performed on error prone PCR libraries (64) may reflect the fact that the required base substitutions are not as common as the base substitutions for previously identified tazobactam resistance mutations. We speculate that we readily identified these mutations because PFunkel provides a less biased and much more comprehensive library of mutations than error prone PCR.

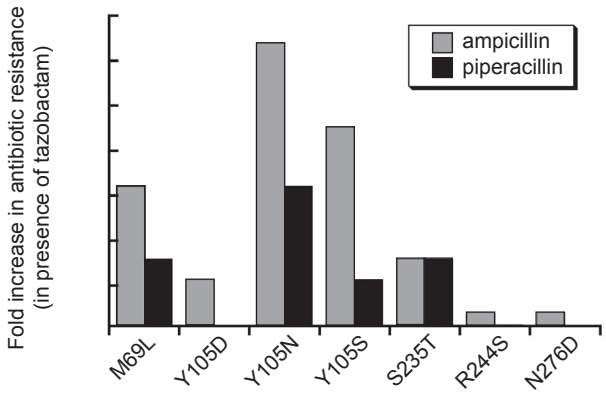


Figure 2.5. Tazobactam resistance of selected alleles**.** The increase in ampicillin or piperacillin resistance is reported as the fold increase (over *TEM-1*) in the minimum inhibitory concentration (MIC) of the antibiotic in the presence of 6 µg/ml tazobactam. MIC assays performed in √2-fold increments of antibiotic concentration. Median MIC values of three replicates were used. Data for all replicates is in Tables 2.8 and 2.9.

Table 2.8. Ampicillin MIC values for selected alleles.

	MIC[a] ampicillin (µg/ml)							
	– tazobactam					+ tazobactam[b]		
Replicate	1	2	3	Median		1	2	3	Median
none	8192	8192	8192	8192		16	22.6	22.6	22.6
M69L	5792	8192	8192	8192		512	724	724	724
Y105D	1448	1448	1448	1448		256	256	362	256
Y105N	4096	5792	5792	5792		1448	1448	1448	1448
Y105S	2896	2896	2896	2896		724	1024	1024	1024
S235T	8192	8192	8192	8192		256	362	512	362
R244S	2896	4096	4096	4096		64	128	90.5	90.5
N276D	8192	8192	8192	8192		90.5	90.5	128	90.5

[a] Median value of three replicates. MIC assays performed in √2-fold increments (Mueller Hinton broth-agar, 10^4 CFU/spot, 37°C for 12 hours).
[b] tazobactam added to 6 µg/ml

Table 2.9. Piperacillin MIC values for selected alleles.

	MIC[a] piperacillin (μg/ml)							
	– tazobactam				+ tazobactam[b]			
Replicate	1	2	3	Median	1	2	3	Median
none	2896	2896	2896	2896	1	1.4	1.4	1.4
M69L	2048	2896	2896	2896	22.6	32	22.6	22.6
Y105D	1.4	1.4	1.4	1.4	1	1	1	1
Y105N	2048	2048	2048	2048	32	45	45	45
Y105S	1024	1448	1448	1448	16	16	16	16
S235T	2896	2896	2896	2896	32	16	22.6	22.6
R244S	1024	1448	1448	1448	2	2.83	2	2
N276D	2896	2896	2896	2896	1	1.4	2	1.4

[a] Median value of three replicates. MIC assays performed in √2-fold increments (Mueller Hinton broth-agar, 10^4 CFU/spot, 37°C for 12 hours).
[b] tazobactam added to 6 μg/ml

2.3.10 PFunkel mutagenesis using a dsDNA template

The methods described above require that the gene targeted for mutation is in a phagemid (a plasmid containing the f1 phage origin) and require the production of phage particles from which the dU-ssDNA template is isolated. Although preparation of such a template is straightforward, we sought to expand the method to be applicable to any plasmid and to simplify the protocol by eliminating the need for phage entirely. Phage-less PFunkel (Figure 2.6, Table 2.2) achieves this by utilizing a dU-dsDNA plasmid template. After the mutation-containing second strand synthesis, UDG and ExoIII are added to degrade both strands of the dU-dsDNA template. The newly-synthesized, circular ssDNA is then converted to dsDNA using the reverse oligonucleotide. Like PFunkel using a ssDNA template, the reaction can be performed in a single tube using a thermocycler.

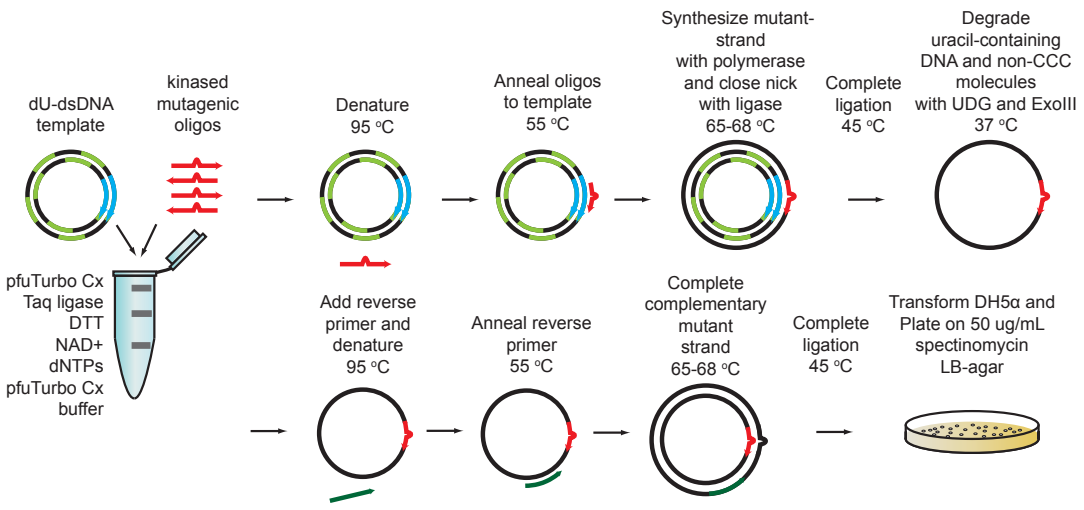


Figure 2.6. Schematic of PFunkel using a dsDNA template. The chief differences from the protocol of Figure 2.1 are the use of a dsDNA template instead of a ssDNA template and the degradation of the dU-containing template before the third strand synthesis.

We tested PFunkel using a dsDNA template for site-directed mutagenesis and multiple-site mutagenesis. For site-directed mutagenesis (using the c542a mutagenic primer), we obtained 708,000 transformants, and 10 of 10 randomly selected colonies had the desired mutation and no undesired mutations. For multiple-site mutagenesis, we attempted to create four specific mutations at distant sites in the gene. We obtained 445,000 transformants. Four of 10 colonies had all four mutations, the remainder either were wildtype (5 colonies) or had less than 4 mutations (1 colony). We speculate that the apparent lower efficiency of multi-site mutagenesis using a dsDNA template may result from the difficulty in getting all four primers to simultaneously anneal to a dsDNA template (as opposed to a ssDNA template) or difficulty in degrading the dsDNA template. Although we have not tested comprehensive codon mutagenesis using a dsDNA template, we believe it is feasible.

2.4 DISCUSSION

PFunkel offers a very efficient method for site-directed mutagenesis at single or multiple-sites simultaneously. However, the real power of PFunkel lies in the ability to make extensive, user defined libraries of single or multiple mutations. PFunkel can be used for alanine scanning mutagenesis (65) to create all possible alanine substitutions, or a user-defined subset thereof in a single reaction. Comprehensive codon mutagenesis using PFunkel efficiently makes libraries for deep mutational scanning (66) without the need for the costly and time-consuming construction of separate libraries for every codon analyzed. Compared to error prone PCR (67), which can practically access only ~30% of the possible amino acid substitutions in a gene, comprehensive codon mutagenesis allows effective access to all 100%. For directed evolution studies, the generation of diversity by comprehensive codon mutagenesis will allow access to unique mutational pathways not readily explored by current methods – enabling the identification of unique proteins with improved properties. PFunkel can efficiently create defined mutagenic diversity at multiple sites simultaneously and is thus tailor-made for the creation of computationally designed libraries for subsequent screening or selection strategies.

2.5 ACKNOWLEDGEMENTS

We thank Tim Whitehead for helpful comments on the manuscript.

CHAPTER 3

THE GENETIC CODE CONSTRAINS YET FACILITATES DARWINIAN EVOLUTION

SUMMARY

An important goal of evolutionary biology is to understand the constraints that shape the dynamics and outcomes of evolution. Here we address the extent to which the structure of the standard genetic code constrains evolution by analyzing adaptive mutations of the antibiotic resistance gene *TEM-1* β-lactamase and the fitness distribution of codon substitutions in two influenza hemagglutinin inhibitor genes. We find that the architecture of the genetic code significantly constrains the adaptive exploration of sequence space. However, the constraints endow the code with two advantages: the ability to restrict access to amino acid mutations with a strong negative effect and, most remarkably, the ability to enrich for adaptive mutations. Our findings support the hypothesis that the standard genetic code was shaped by selective pressure to minimize the deleterious effects of mutation yet facilitate the evolution of proteins through imposing an adaptive mutation bias.

3.1 INTRODUCTION

The genetic code plays a central role in evolutionary processes defining the relationship between DNA and protein sequences. The genetic code limits the mutational exploration of sequence space (1), since single base changes in codons can access only about six of the nineteen possible amino acid substitutions and simultaneous multiple-base changes in a codon are rare (2). Furthermore, the genetic code is biased towards conservative amino acid mutations (3). As a result, most mutational trajectories have a low probability, and probable mutational trajectories tend to be conservative in nature. Thus, very similar genes may follow different evolutionary trajectories in part because the genes' mutational neighborhoods are different (i.e. the likely amino acid substitutions are different due to the two genes having different but synonymous codons) (68-70). What has not been experimentally addressed is the extent to which the evolution of a single gene is constrained or facilitated by the architecture of the genetic code. For a particular evolutionary outcome, how many superior fitness peaks are nearby that could have been reached if only the genetic code was arranged differently?

To the extent that the genetic code restricts a gene from evolving to higher fitness peaks, one may wonder about the possible benefits of alternative codes. However, the standard genetic code's organization makes it apparent that the relationship between DNA triplets and amino acids was not arrived at randomly. The code's arrangement has been proposed to result from the inherent interactions between amino acids and their cognate nucleotide triplets (5), the biochemical pathways through which amino acids are synthesized (10), selective pressure to minimize the deleterious effects of mutations and mistranslations (3,7), and the difficulty in changing the code once it is established (9).

These basic theories have been further developed primarily through theoretical and simulation approaches (6). However, mutational bias, such as that arising from how the code is arranged or the nature of spontaneous mutations, may shape evolution (71,72). Thus, the architecture of the genetic code may facilitate the evolution of genes and proteins (73). For example, a code arranged to make adaptive mutations more likely would have provided an adaptive advantage over one that did not early in evolution when there may have been competing genetic codes. If so, the standard genetic code might still exhibit this property today. Does the standard genetic code enrich for adaptive mutations?

Here we experimentally address the extent to which the genetic code restricts access to beneficial alleles and whether the code's constraints provide advantages for the evolution of proteins. We find that although the code's architecture significantly limits evolutionary outcomes, it minimizes the deleterious cost of mutation and enriches for beneficial mutations – two properties that facilitate evolution.

3.2 MATERIALS AND METHODS

3.2.1 *TEM-1* β-lactamase libraries and constructs

All libraries and variants of the *TEM-1* gene were created using PFunkel mutagenesis as previously described (74).

3.2.2 Library selections

All antibiotics and chemical reagents for selections and MIC assays were obtained from Sigma-Aldrich. For the selection of alleles conferring cefotaxime resistance equivalent to that of *GKTS*, a previously described library (74) in which the codons for

A42, E104, M182, and G238 were randomized (NNN) was plated on LB-agar plates containing 50 μg/mL spectinomycin, 50 μM IPTG, and 8 μg/mL or 16 μg/mL cefotaxime. The library was plated at a cell density of 1,900 or 19,000 CFU/cm^2 on the 8 μg/mL cefotaxime plates and 190,000 CFU/cm^2 on the 16 μg/mL cefotaxime plate. Plates were incubated at 37°C for 17 hrs. Forty colonies of ~1000 that grew on the 8 μg/mL plate and the largest 10 colonies on the 16 μg/mL plate were selected for individual screening by plate MIC assay. For colonies that passed the screen, plasmid DNA was isolated from overnight cultures using the Qiagen QIAprep Spin Miniprep kit (27106) and the *TEM-1* allele was sequenced. Unique plasmids were retransformed into fresh DH5α *E. coli* cells and the MIC determined by both plate and liquid MIC assays.

For the selection of *TEM-1* alleles with a single amino acid substitution that provides increased resistance to cefotaxime, two previously described *TEM-1* comprehensive codon libraries were used: CCM1 and CCM2 (74). DH5α cells bearing CCM1 were plated at a density of 100-300 CFU/cm^2 on LB-agar plates containing 50 μg/mL spectinomycin, 50 μM IPTG and 0.04 μg/mL cefotaxime. NEB 5-alpha F'I^q cells bearing CCM2 were plated as above except the cell density was 150-600 CFU/cm^2, the cefotaxime was 0.02 μg/mL, and the IPTG was 300 μM. The concentrations of cefotaxime used correspond to the MIC conferred by *TEM-1* in DH5α and NEB 5-alpha F'I^q cells. Plates were incubated at 37 °C for 17 hrs. The *TEM-1* gene of randomly selected colonies was sequenced. Since any particular amino acid substitution is relatively rare in the libraries, we used the criteria that an amino acid substitution had to be observed twice for us to categorize it as adaptive.

3.2.3 MIC assays

For MIC assay on agar plates, cultures of variants were prepared in LB broth at 37 °C with shaking at 250 rpm until all cultures reached saturation, ~24 hrs. Cultures were diluted 100-fold in LB broth, and incubated for ~2.5 hrs until the OD was about 0.3. The cultures were diluted to 10^4 CFU/μL and 1 μL spotted on Mueller-Hinton agar plates containing 50 μg/mL spectinomycin, 50 μM IPTG, and √2-fold increasing concentrations of cefotaxime. The plates were incubated at 35 °C for 20 hrs. The MIC was determined as the minimal concentration at which no growth was observed.

For liquid MIC assays, the initial cultures were prepared as above and then diluted to a concentration of $1x10^6$ CFU/μL in Mueller-Hinton broth. A total of 150 μL of this diluted culture was added to wells of a 96-well assay plate along with 150 μL of Mueller-Hinton broth containing 100 μg/mL spectinomycin, 100 μM IPTG, and 2-fold increasing concentrations of cefotaxime. The plate was covered and sealed in a plastic bag and incubated at 35 °C for 20 hrs. The MIC was determined as the minimal concentration at which no visible growth was observed.

The above two MIC tests used different temperatures and media than the selections. The MIC assay conditions used are standardized conditions for quantifying β-lactam resistance that allow comparisons to other studies (75).

3.2.4 Enrichment values of experimentally observed codon substitutions

Enrichment values for each codon substitution introduced in *HB36.4* and *HB80.3* were determined by using custom Matlab scripts to analyze the Illumina deep sequencing data on the libraries before and after selection for hemagluttinin binding. The data was filtered as in the study by Whitehead et al. (20) to include sequencing reads with only one

codon substitution, and the final list to include only codon substitutions with at least 100 sequencing counts in the reference library. The enrichment value *E* was calculated as

$$E = \log_2\left(\frac{(selected\ library\ counts)/(total\ selected\ library\ counts)}{(reference\ library\ counts)/(total\ reference\ library\ counts)}\right) \quad \text{(Equation 1)}$$

Thus, *E* quantifies the relative prevalence of an allele in the selected library compared to the reference library (the naïve library). The enrichment value of the wildtype sequence was determined by averaging the enrichment values of all codons synonymous with the wildtype. Codon substitution counts, enrichment values, and the wildtype enrichment values were consistent with values for amino acid substitutions presented by Whitehead et al. (20).

3.2.5 The genetic code's enrichment of adaptive mutations and meta-analysis

The percent enrichment and p-values were determined as described in Table 3.8. Meta-analysis on the p-values was performed using the Stouffer's Z-trend method (weighted Z score) using the program MetaP (http://people.genome.duke.edu/~dg48/metap.php).

3.3 RESULTS

3.3.1 The natural and in vitro evolution of TEM-1 β-lactamase for conferring cefotaxime resistance converges on the same set of mutations

We chose to examine the genetic code's constraints on evolution with the antibiotic resistance *TEM-1* gene encoding TEM-1 β-lactamase – a gene that has provided many insights into how epistasis constrains evolution (75-78). TEM-1 hydrolytically inactivates β-lactam drugs such as penicillin, but has very low activity on the third generation cephalosporin β-lactam cefotaxime. Clinically isolated alleles of

TEM-1 conferring elevated antibiotic resistance arise through accumulation of point mutations (i.e. 1-bp substitutions). For example, *TEM-52* differs from *TEM-1* by three point mutations resulting in the E104K/M182T/G238S mutations (79) that increase cefotaxime resistance ~4,000-fold (75). The in vitro evolution of *TEM-1* mimics its natural evolution (80). Six independent in vitro evolution studies that applied selective pressure for increased cefotaxime resistance found the E104K/M182T/G238S combination of mutations in the best alleles (78,80-84). A fourth mutation (A42G) (83) that also arises from a point mutation increases cefotaxime resistance to about 33,000-fold over *TEM-1* (75). The high fitness of the gene bearing the A42G/E104K/M182T/G238S mutations (referred to here as *GKTS*) has not been surpassed by increasing the mutation rate (78,84), using mutator strains of bacteria (82), forcing explorations of alternative trajectories through use of bottlenecks (78), or employing computational protein optimization strategies (which are not constrained by the genetic code) (85). This suggests that the evolutionary outcome of *GKTS* is largely reproducible and inevitable, given a strong selective pressure for cefotaxime resistance (75). Among the accessible local optima for cefotaxime resistance on the β-lactamase fitness landscape, *GKTS* may be the global optimum.

To what extent did the architecture of the genetic code direct this outcome? There are 20^4 -1 = 159,999 possible amino acid combinations at these four positions in TEM-1 (including combinations with up to three wildtype amino acids). However, only 2743 (i.e. 7x8x7x7 - 1) or 1.7% of these are readily accessible combinations since they do not require simultaneous multiple mutations in any one codon, which is a rare occurrence. Synonymous mutations followed by a second point mutation can expand the accessible

amino acid combinations, but only to some extent. Also, subsequent second mutations in codons with previously accumulated beneficial mutations can occur, but this requires that both mutations be beneficial and that the second mutation increases the fitness of the gene. This significant constraint will keep the readily accessible combination of mutations at these four codons low. More to the point, such double mutations are not present in the best cefotaxime resistance alleles arrived at by natural or in vitro evolution of *TEM-1*. The rarity of natural adaptive mutations with multiple base substitutions in a single codon is exemplified by a recent study that examined 516 spontaneous ceftazidime resistant isolates of *Burkholderia thailandensis* and found 29 different codon substitutions in the *penA* β-lactamase gene that provided this resistance – all of which were point mutations (86). In addition, the occurrence of reciprocal sign epistasis will further constrain which combination of amino acids are accessible by evolution (87).

3.3.2 The genetic code constrains the evolution of *TEM-1*

To test the extent to which the genetic code constrains the evolution of *TEM-1*, we asked whether there exist other amino acid combinations at these four positions that provide fitness equal to or better than *GKTS*. We employed a library in which these four codons were randomized at all three base positions (74). We placed the mutated gene downstream from the IPTG-inducible *tac* promoter on a plasmid with the *p15A* origin (copy number ~10), as in previous in vitro evolution experiments with *TEM-1* (78,80). Our library consisted of 5.8 million transformants (short of the theoretical 4^{12} = 16.8 million DNA variants, but in excess of the possible 20^4 = 160,000 protein variants), and the majority of library members contained mutations at all four positions (74). Although we subsequently determined in separate experiments that the degenerate oligos used to

make the library were enriched for G's by about 2.2-fold (74), a large fraction of the protein variants are likely to be present in the library.

We subjected this library to selections for cefotaxime resistance equivalent or superior to that achieved by *GKTS*. We performed a secondary screen on 50 of the resulting ~1000 colonies for those with a MIC at or above that conferred by *GKTS*. Clones passing this screen were sequenced. The plasmid DNA from unique clones was retransformed into fresh DH5α *E. coli* and the cefotaxime MIC determined by solid and liquid media growth assays.

The sequences and corresponding MICs revealed that there are many alleles with equivalent or superior combinations of amino acids at these four positions (Table 3.1). Of the 17 identified alleles (11 unique amino acid combinations), only four were identified more than once, indicating that there are additional high fitness alleles yet to be found. Although we observe small differences (up to two-fold) between synonymous alleles in the plate MIC assay (Table 3.1), this assay showed variability in replicate experiments of up to two-fold in some instances (Table 3.2). Six unique amino acid sequences differed from GKTS at two of the four positions. Most strikingly, 55% (6 of 11) of the amino acid combinations identified require more than one point mutation in at least one codon (i.e. Hamming distance >4). We find the existence of numerous equivalent or superior alleles nearby what appeared to be a dominant resistance allele quite striking. This experiment shows that there are many alleles equivalent or superior to *GKTS* nearby in sequence space that are not readily accessed by natural or in vitro evolution.

Table 3.1. Cefotaxime resistance of selected TEM-1 β-lactamase alleles.

Colony [a]	For positions 42-104-182-238		# base changes in each codon [c]	MIC (μg/mL) [d]	H (actual) [e]	H (minimum) [f]
	Amino acids [b]	Codons				
-	No TEM-1 gene			0.08	-	-
TEM-1	A-E-M-G	gca-gag-atg-ggt	0-0-0-0	0.08	0	0
GKTS	G-K-T-S	gga-aag-acg-agt	1-1-1-1	90.5	4	4
43, 48	G-K-**M**-**A**	ggg-aag-atg-gcg	2-1-0-2	90.5	5	3
24	G-K-**M**-S	ggg-aag-atg-tca	2-1-0-3	45.3	6	3
2	G-K-**K**-**A**	ggg-aag-aag-gct	2-1-1-1	64	5	4
6	G-K-T-**A**	gga-aag-acg-gct	1-1-1-1	90.5	4	4
34		ggg-aag-acg-gcg	2-1-1-2	181	6	4
9		ggg-aag-aca-gcc	2-1-2-2	181	7	4
32	G-K-T-S	ggg-aag-acg-tcg	2-1-1-3	90.5	7	4
1	G-K-**A**-**A**	ggg-aag-gcg-gct	2-1-2-1	90.5	6	5
38	G-K-**A**-S	ggg-aag-gcg-agc	2-1-2-2	90.5	7	5
16		ggg-aag-gcc-agc	2-1-3-2	64	8	5
14	G-K-**Q**-**A**	ggg-aag-cag-gca	2-1-2-2	128	7	5
5, 31		ggg-aag-cag-gcc	2-1-2-2	90.5	7	5
7, 15		ggc-aag-caa-gca	2-1-3-2	64	8	5
46	G-K-**S**-**A**	ggg-aag-agc-gct	2-1-2-1	128	6	5
33	G-K-**S**-S	ggg-aaa-agt-agt	2-2-2-1	90.5	7	5
3	G-**R**-**S**-S	ggg-cgg-agc-tcg	2-2-2-3	64	9	6
45, 49		ggt-aga-tct-tcg	2-3-3-3	128	11	6

[a]Two numbers indicates that the allele was found twice.

[b]Bold indicates amino acids differing from those in GKTS

[c]Relative to *TEM-1*.

[d]Median value of three replicates. Assays performed in √2 increments of cefotaxime (Mueller-Hinton-agar, 10^4 CFU/spot, 35°C for 20 hours). Data for all replicates are in Table 3.2. MICs determined by Mueller-Hinton broth liquid growth assay at 35 °C can be found in Table 3.3.

[e]Hamming distance between the allele and *TEM-1*.

[f]Minimum Hamming distance to achieve same amino acid sequence

Table 3.2. Replicate cefotaxime resistance of *TEM-1* alleles by plate assay

Colony	For positions 42-104-182-238		MIC[a](µg/ml)			
	Amino acids	Codons	Replicate 1	Replicate 2	Replicate 3	Median
-	No TEM-1 gene		0.08	0.08	0.08	0.08
TEM-1	A-E-M-G	gca-gag-atg-ggt	0.08	0.08	0.08	0.08
GKTS	G-K-T-S	gga-aag-acg-agt	90.5	90.5	90.5	90.5
43, 48	G-K-M-A	ggg-aag-atg-gcg	90.5	90.5	90.5	90.5
24	G-K-M-S	ggg-aag-atg-tca	45.3	45.3	64	45.3
2	G-K-K-A	ggg-aag-aag-gct	64	64	128	64
6	G-K-T-A	gga-aag-acg-gct	90.5	90.5	128	90.5
34		ggg-aag-acg-gcg	181	181	181	181
9		ggg-aag-aca-gcc	128	181	181	181
32	G-K-T-S	ggg-aag-acg-tcg	90.5	64	128	90.5
1	G-K-A-A	ggg-aag-gcg-gct	90.5	90.5	128	90.5
38	G-K-A-S	ggg-aag-gcg-agc	64	64	90.5	64
16		ggg-aag-gcc-agc	90.5	90.5	90.5	90.5
14	G-K-Q-A	ggg-aag-cag-gca	90.5	90.5	128	90.5
5, 31		ggg-aag-cag-gcc	64	64	64	64
7, 15		ggc-aag-caa-gca	90.5	128	128	128
46	G-K-S-A	ggg-aag-agc-gct	181	128	90.5	128
33	G-K-S-S	ggg-aaa-agt-agt	90.5	64	128	90.5
3	G-R-S-S	ggg-cgg-agc-tcg	64	64	90.5	64
45, 49		ggt-aga-tct-tcg	181	128	128	128

[a] Plate MIC assays performed in $\sqrt{2}$ increments of cefotaxime (Mueller-Hinton-agar, 10^4 CFU/spot, 35°C for 20 hours).

Table 3.3. Cefotaxime resistance of *TEM-1* alleles by liquid assay

Colony	For positions 42-104-182-238		MIC[a](μg/ml)			
	Amino acids	Codons	Replicate 1	Replicate 2	Replicate 3	Median
-	No TEM-1 gene		0.08	0.08	0.08	0.08
TEM-1	A-E-M-G	gca-gag-atg-ggt	0.08	0.08	0.08	0.08
GKTS	G-K-T-S	gga-aag-acg-agt	2048	4096	2048	2048
43, 48	G-K-M-A	ggg-aag-atg-gcg	2048	2048	2048	2048
24	G-K-M-S	ggg-aag-atg-tca	2048	2048	2048	2048
2	G-K-K-A	ggg-aag-aag-gct	2048	2048	2048	2048
6	G-K-T-A	gga-aag-acg-gct	2048	2048	2048	2048
34		ggg-aag-acg-gcg	2048	2048	2048	2048
9		ggg-aag-aca-gcc	2048	2048	2048	2048
32	G-K-T-S	ggg-aag-acg-tcg	4096	4096	4096	4096
1	G-K-A-A	ggg-aag-gcg-gct	2048	2048	2048	2048
38	G-K-A-S	ggg-aag-gcg-agc	4096	2048	2048	2048
16		ggg-aag-gcc-agc	2048	2048	2048	2048
14	G-K-Q-A	ggg-aag-cag-gca	2048	2048	2048	2048
5, 31		ggg-aag-cag-gcc	2048	2048	2048	2048
7, 15		ggc-aag-caa-gca	2048	2048	2048	2048
46	G-K-S-A	ggg-aag-agc-gct	2048	2048	2048	2048
33	G-K-S-S	ggg-aaa-agt-agt	2048	2048	2048	2048
3	G-R-S-S	ggg-cgg-agc-tcg	2048	2048	2048	2048
45, 49		ggt-aga-tct-tcg	2048	2048	2048	2048

[a] Liquid MIC assays performed in 2-fold cefotaxime increments (Mueller-Hinton broth, $5x10^5$ CFU/culture, 35°C for 20 hours)

To address whether there is a mutational pathway to any of the alleles with a Hamming distance >4, we chose *GKQA* (codons: ggg-aag-cag-gca) as a representative allele and constructed the 14 combinations of these four codon substitutions. We considered each codon substitution (whether a point mutation or a multi-bp substitution) as a single mutational step in order to ask whether *GKQA* could be reached if the genetic code were arranged differently such that each of the required amino acid substitutions were possible with a point mutation. We tested the cefotaxime resistance of these variants and assessed the feasibility of the 24 possible trajectories from *TEM-1* to *GKQA*.

We assumed that the evolution of *TEM-1* fits the strong selection/weak mutation model of evolution by which the time to fixation or loss of a mutation is much shorter than the time between mutations. Thus, we required that mutations accumulate one at a time with increasing fitness at each step for a trajectory to be deemed feasible, as in a previous study (75). Nine of the 24 possible trajectories were feasible (Figure 3.1). Four trajectories ended at an intermediate (*GKMA*) with equivalent resistance to *GKQA*. Like the feasible trajectories for evolving *GKTS* (75), the first mutation necessarily occurs at positions 104 or 238, and the fittest double mutant has mutations at both positions (the difference is that 238 is mutated to A instead of S for *GKQA*). A lack of a mutational trajectory cannot explain why *GKQA* has not been found in the natural or in vitro evolution of *TEM-1*. Instead, we posit that the requirement for multiple mutations in a single codon is one reason that makes this allele's occurrence unlikely. Thus, the architecture of the genetic code constrains evolution by making some viable mutational trajectories improbable.

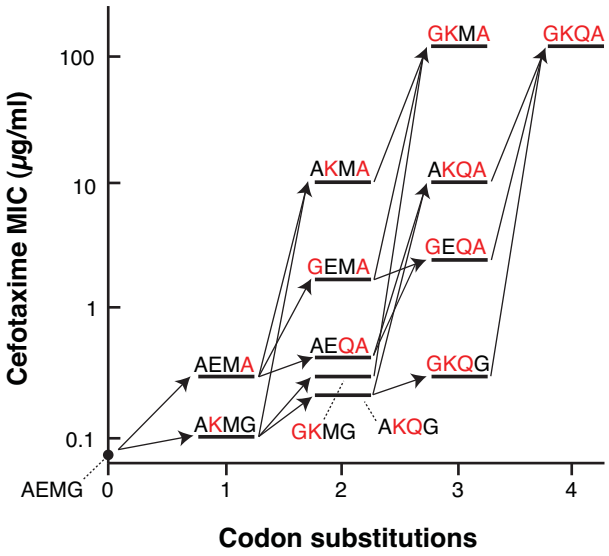


Figure 3.1. Feasible trajectories for evolving *GKQA* (colony 14) from *TEM-1* (i.e. *AEMG*) by accumulation of codon substitutions one at a time. Mutations are shown in red. Of the 24 possible trajectories, five end with *GKQA* and four end with *GKMA*, an allele with equivalent fitness to *GKQA*. Cefotaxime resistance was measured by plate assay as in Table 3.1 and the value reported represents the median of three replicates. Data for all replicates is provided in Table 3.4.

Table 3.4. Replicate cefotaxime resistance of *TEM-1* alleles for Figure 3.1.

Allele	Codon Substitutions	Cefotaxime MIC[a] (μg/ml)			
		Replicate 1	Replicate 2	Replicate 3	Median
AEMG	0	0.06	0.08	0.08	0.08
GEMG	1	0.08	0.06	0.08	0.08
AKMG	1	0.08	0.11	0.16	0.11
AEQG	1	0.08	0.08	0.08	0.08
AEMA	1	0.32	0.32	0.32	0.32
GKMG	2	0.32	0.32	0.32	0.32
GEQG	2	0.06	0.08	0.08	0.08
GEMA	2	1.81	1.81	1.81	1.81
AKQG	2	0.11	0.23	0.23	0.23
AKMA	2	10.24	10.24	10.24	10.24
AEQA	2	0.32	0.45	0.45	0.45
GKQG	3	0.32	0.32	0.32	0.32
GKMA	3	115.85	115.85	115.85	115.85
GEQA	3	1.81	2.56	2.56	2.56
AKQA	3	10.24	10.24	10.24	10.24
GKQA	4	115.85	115.85	115.85	115.85

[a]MIC assays performed in $\sqrt{2}$ increments of cefotaxime (Mueller-Hinton-agar, 10^4 CFU/spot, 35°C for 20 hours).

3.3.3 Epistasis and mutational bias also constrain the exploration of sequence space.

Alleles such as *GKMA* and *GKTA* provide resistance equivalent to *GKTS* (Table 3.1) and have amino acid combinations that can be reached by three or four point mutations, respectively, yet these combinations have not been previously identified in laboratory evolution experiments. We speculate that there are two reasons why these alleles have not been previously identified. The first reason is epistasis. For example, G238A was more common in the identified alleles than G238S and is present in both *GKMA* and *GKTA*. However, when occurring as the only mutation in TEM-1, G238S provides a 4-fold higher k_{cat}/K_m and a 4-fold higher MIC than G238A (88). Since G238S is the amino acid substitution that, by itself, increases cefotaxime resistance the most, it is most likely to be fixed first. The apparent equivalence of G238A and G238S among our

selected alleles is illustrative of the epistatic nature of mutations. The second reason is mutational bias. For example, the G:C → C:G mutation necessary for G238A is considerably less common than the G:C → A:T mutation for G238S in error-prone PCR reactions (89) and spontaneously in *E. coli* (90).

3.3.4 The genetic code minimizes the fitness cost of mutations

There is no arrangement of a 20 amino acid / 64 codon genetic code that would not significantly limit the types of amino acid substitutions that are readily accessible, since a single codon can be mutated only to nine other codons with a point mutation. Thus, different genetic codes will constrain evolutionary paths in different ways. The adaptive theory of the origin of the genetic code postulates that the code's conservative architecture is a result of selective pressure to minimize the deleterious effects of point mutations and mistranslation errors (3,7). The adaptive theory predicts that for non-synonymous mutations the average fitness cost of point mutations should be less than that of 2-bp and 3-bp substitutions. However, such a test of the adaptive theory (and of the conservative nature of the code) has never been systematically applied to any gene. How does the mutational fitness distribution partition between 1-, 2- and 3-bp codon substitutions of a gene?

To address this question we examined the distribution of fitness effects of 1896 unique single amino acid substitutions in two genes that were previously modified through a combination of computational design and directed evolution to inhibit H1N1 influenza hemagglutinin (20,91). Inhibitor HB36.4 derives from Apc36109 from *Bacillus stearothermophilus* and HB80.3 from the Myb domain of the Rad transcription factor from *Antirrhinum majus* (91). Since the natural proteins are not inhibitors of

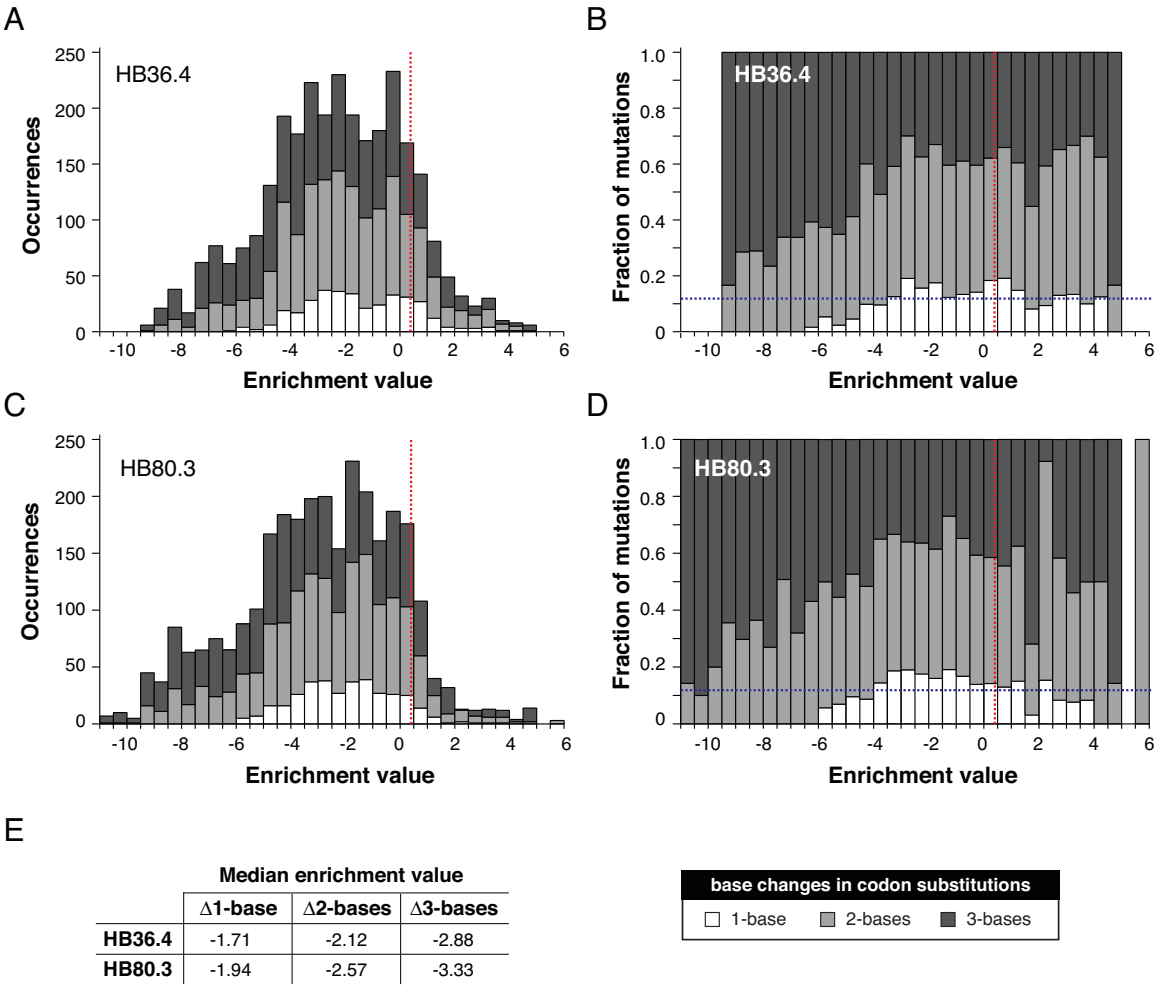
hemagglutinin, we suggest that HB36.4 and HB80.3 should be viewed as genes that are not evolutionarily mature for hemagglutinin inhibition. Whitehead et al. (20) created NNK degenerate codon libraries consisting of all possible single amino acid substitutions in all 51 positions in HB80.3 and 53 of 93 positions of HB36.4. In NNK libraries, the first two nucleotides in a codon can be any base, but the third nucleotide is limited to G or T to reduce the frequency of nonsense codons while still allowing all possible amino acids. The libraries were subjected to deep sequencing before and after selection for hemagglutinin binding in a yeast display format. In their study, the base 2 logarithm of the ratio of the frequencies of each amino acid substitution in the selected versus unselected libraries – referred to as the enrichment value – served as a proxy for the change in free energy of binding. Several lines of evidence support the suitability of this proxy (20).

Since differences in enrichment values for synonymous mutations were considerably smaller than differences between non-synonymous mutations (20), we assigned the enrichment values of each amino acid substitution to its respective codon substitutions and used this value as a proxy for the change in gene fitness caused by the 5857 codon substitutions (i.e. the number of unique codon substitutions that can code for the 1896 unique amino acid substitutions observed). The distribution of these fitness effects (Figure 3.2A-D) indicates that the standard genetic code's architecture asymmetrically partitions fitness effects of amino acid substitutions between point mutations and multi-bp codon substitutions and minimizes the fitness cost of point mutations. The average cost of mutation for point mutations is substantially less than for 2-bp substitutions, and 3-bp substitutions have the highest average fitness cost (Figure

3.2E). Mutations causing a smaller decrease in fitness are enriched in point mutations and mutations with the largest negative effect are almost exclusively 2- and 3-bp changes. The same distribution trends are observed when we determined the enrichment values for the 2813 experimentally-observed codon substitutions (Figure 3.3). Among beneficial mutations, the median effect of mutations for multi-bp mutations was marginally higher than that of point mutations (Table 3.5).

Table 3.5. Median enrichment value of adaptive mutations as a function of the number of base changes in the codon.

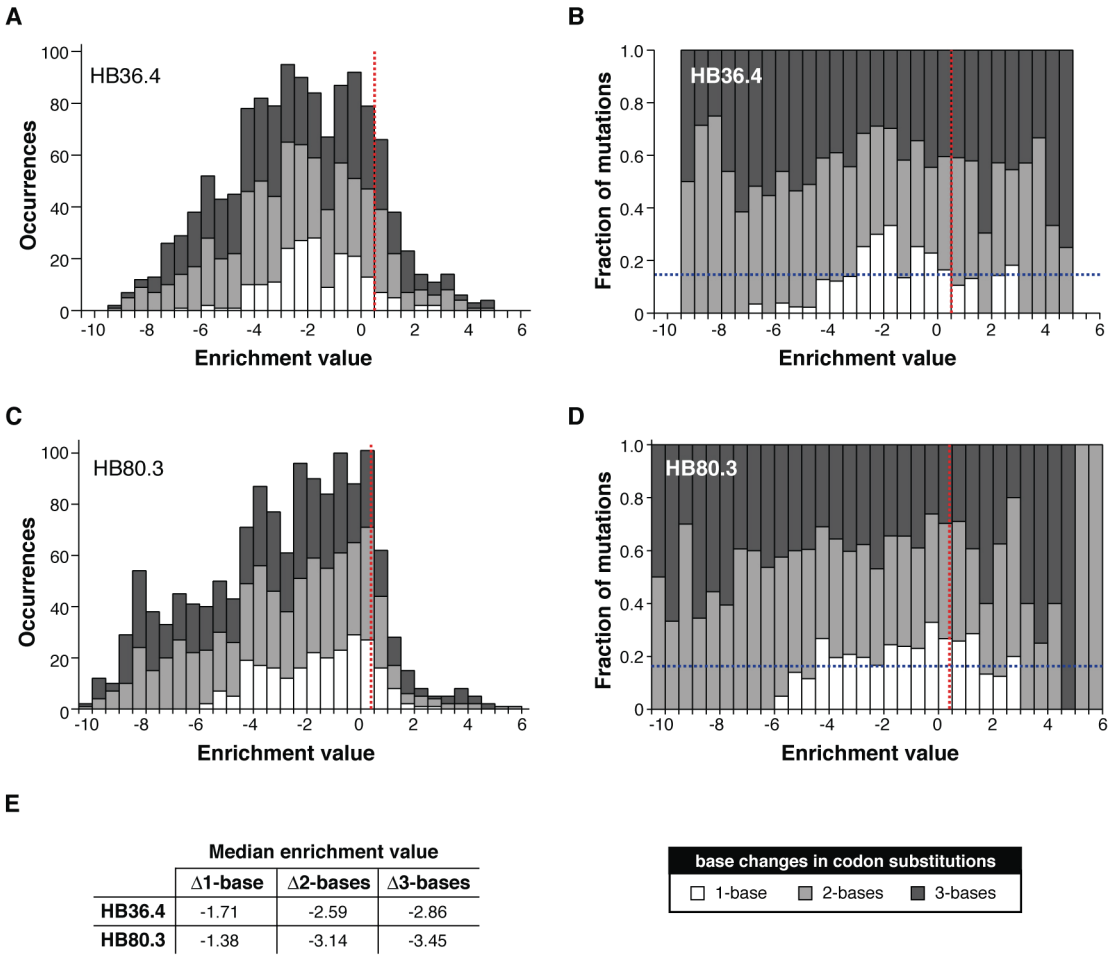
	Median enrichment value		
Gene	Δ1-base	Δ2-bases	Δ3-bases
HB36.4	0.933	1.172	1.192
HB80.3	0.956	1.025	1.285



Median enrichment value

	Δ1-base	Δ2-bases	Δ3-bases
HB36.4	-1.71	-2.12	-2.88
HB80.3	-1.94	-2.57	-3.33

Figure 3.2. Distribution of fitness effects of non-synonymous codon substitutions in (**A**, **B**) *HB36.4* and (**C**, **D**) *HB80.3*. The distribution is partitioned into codon changes with 1-, 2-, and 3-base changes. The red dashed vertical line indicates the enrichment value of the parental genes, and the blue dashed horizontal bar indicates the fraction of all possible mutations of the gene that are point mutations. Enrichment values for parental genes are slightly greater than zero because most mutations have a negative effect on fitness. (**E**) Median enrichment values for types of codon substitutions. Distributions based on codon enrichment values instead of amino acid enrichment values is provided in Figure 3.3.



Median enrichment value

	Δ1-base	Δ2-bases	Δ3-bases
HB36.4	-1.71	-2.59	-2.86
HB80.3	-1.38	-3.14	-3.45

Figure 3.3. Distribution of fitness effects of codon substitutions in (**A**, **B**) *HB36.4* and (**D**, **E**) *HB80.3* determined using codon enrichment values for experimentally observed codon substitutions. The distribution is partitioned into codon changes with 1-, 2-, and 3-base changes. The red dashed vertical line indicates the enrichment value of the parental genes, and the blue dashed horizontal bar indicates the fraction of all possible mutations of the gene that are point mutations. The difference between Fig. 2 and this figure is that in Fig. 2, the amino acid enrichment values from Whitehead et al. (20) were used and all the synonymous codon substitutions that could encode an amino acid substitution (whether they were present in the library or not) were assigned the same enrichment value. In this figure, only experimentally observed codon substitutions with >100 counts in the unselected library were included, and the enrichment values were calculated on an individual codon basis. (**E**) Median enrichment values for types of codon substitutions.

We interpret these results (Figure 3.3E) as evidence that the code's arrangement minimizes the fitness cost of amino acid substitutions. An alternative explanation is that the genes are products of evolution under the standard genetic code, and thus their make-

up is such that point mutations will cause minimal deleterious effects. However, this viewpoint is mitigated somewhat by the fact that HB36.4 and HB80.3 were not evolved by nature for hemagglutinin binding, but rather are a product of codon optimization for yeast expression, computational design, and limited laboratory evolution. The genes may better represent ones in the process of evolving rather than "evolutionarily mature" genes.

3.3.5 The genetic code is biased towards adaptive mutations

Figure 3.2 shows that readily accessible amino acid substitutions (i.e. those from point mutations) have smaller deleterious effects on fitness. How does this affect the evolution of proteins? For a gene with L codons there are $19L$ possible amino acid substitutions. The standard genetic code imposes a restriction on which of these $19L$ are likely (i.e. on average, only about $6L$ occur with a point mutation; the exact number for a particular gene will depend on the gene's DNA sequence). In other words, the genetic code provides a set of codon-based rules governing which amino acid substitutions readily occur. Are these rules biased towards adaptive mutations? If so, then the code's arrangement would provide an advantage since accessible amino acid substitutions would be more likely to be adaptive than randomly chosen amino acid substitutions (Figure 3.4).

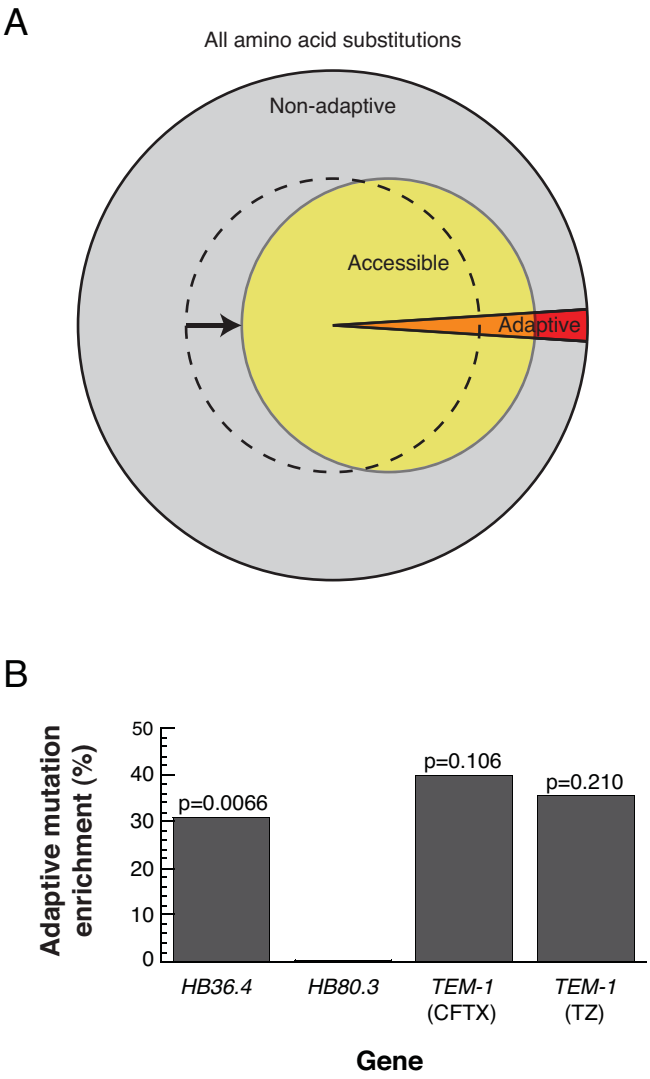


Figure 3.4. Enrichment of adaptive amino acid substitutions of genes by the standard genetic code. The yellow circle depicts a code in which point mutations preferentially access adaptive amino acid substitutions while the dotted circle depicts a non-enriching code that randomly samples amino acid substitutions.

We examined this question in *HB36.4*, *HB80.3*, and *TEM-1* by first identifying an extensive set of adaptive amino acid substitutions in these genes. For our analysis of *HB36.4* and *HB80.3* we utilized the set of amino acid substitutions enriched over wildtype as determined by Whitehead et al. (20). For *TEM-1*, we previously constructed comprehensive codon mutagenesis libraries that consisted of ≥97% of all 18,081 possible single codon substitutions in *TEM-1* (i.e. 63 possible codons x 287 positions in *TEM-1*) (74). From these libraries we previously identified 38 tazobactam resistance alleles (19 unique amino acid substitutions) – tazobactam being an inhibitor of TEM-1 (74). Here,

in an analogous manner, we identified 77 cefotaxime resistance alleles (30 unique amino acid substitutions) by sequencing the *TEM-1* gene of 500 colonies that formed when the library was challenged to grow on plates with elevated levels of cefotaxime (Table 3.6). Whereas the adaptive amino acid substitutions identified in *HB36.4* and *HB80.3* span the range from smallest to largest beneficial effect, those identified in *TEM-1* are the amino acid substitutions with the largest adaptive effect (see section 3.3.6).

Table 3.6. Experimentally identified adaptive codon substitutions for cefotaxime resistance in *TEM-1*

Ambler position	codon		amino acid		occurrences
	WT	mutated	WT	mutated	
69	atg	tgc	M	C	1
		tgt			1
104	gag	aag	E	K	12
		atg		M	2
		cgg		R	3
164	cgt	gca	R	A	1
		gcg			1
		gct			1
		gac		D	1
		gat			3
		ggc		G	1
		ggg			5
		ggt			5
		cac		H	1
		cat			17
		aac		N	7
		agc		S	1
		agt			11
		tca			2
		tcg			14
		tct			3
166	gaa	ccc	E	P	1
		ccg			1
171	gaa	gta	E	V	2
		gtt			1
		tat		Y	2
172	gcc	cac	A	H	3
		cat			4
		cca		P	1

		ccc			2
		cct			2
		aca		T	1
		acc			5
		tac		Y	4
		tat			1
238	ggt	gca	G	A	1
		gcc			2
		gcg			4
		gct			3
		gac		D	1
		gat			16
		gaa		E	2
		gag			3
		aac		N	1
		aat			1
		agc		S	2
		agt			45
		tca			2
		tcc			1
		tcg			3
		tct			2
		aca		T	2
		act			1
240	gag	gca	E	A	1
		gcg			3
		gct			1
		gga		G	8
		ggc			9
		ggg			20
		ggt			10
		cca		P	1
		ccg			8
		cct			2
		agc		S	5
		agt			1
		tca			2
		tcc			1
		tcg			1
		tct			1
		aca		T	1
		act			1
241	cgt	cca	R	P	15
		ccc			13
		ccg			25

		cct			13
243	tct	ggc	S	G	3
		ggg			1

We then calculated the fraction of adaptive amino acid mutations that could be reached with a point mutation and compared that to the fraction of all amino acid substitutions that are possible with a point mutation. For these genes, the standard genetic code enriched for adaptive amino acid substitutions up to 40% (Table 3.7). We assessed the significance of this result by comparing the experimentally determined number of adaptive mutations accessible by a point mutation to the distribution of that value expected if adaptive mutations were chosen at random from all possible mutations (a hypergeometric distribution). The p-value obtained reflects the probability of arriving at that enrichment value or higher by chance under the null hypothesis that adaptive amino acid substitutions are no more likely to be accessible by a point mutation than are all possible amino acid substitutions. Considering the p-values of all four experiments collectively by meta-analysis (92), we find that the enrichment observed in our experiments is significant ($p=0.0027$). This result supports our hypothesis that the standard genetic code, by its limitations on which amino acid substitutions are accessible by a point mutation, facilitates the evolution of proteins by enriching for adaptive mutations.

Table 3.7. Enrichment for adaptive mutations provided by the standard genetic code

Gene	Adaptive advantage	% enrichment of adaptive amino acids[a]
TEM-1	cefotaxime resistance	39.6 (p=0.106)
TEM-1	tazobactam resistance	35.6 (p=0.210)
HB36.4	hemagglutinin binding	30.6 (p=0.0066)
HB80.3	hemagglutinin binding	0.51 (not significant)

[a]Details on this calculation provided in Table 3.8. The p-values provide the probabilities that the observed enrichment was arrived at by chance under the null hypothesis that adaptive mutations are as likely to be accessible by point mutations as non-adaptive mutations.

Table 3.8. Enrichment of adaptive amino acid substitutions by the genetic code.

Gene	Adaptive advantage	Number of amino acid substitutions in protein		Number of identified adaptive amino acids substitutions in protein		Enrichment of adaptive amino acids[a] (%)	p-value[b]
		All	Accessible with a 1-bp substitution	All	Accessible with a 1-bp substitution		
TEM-1	cefotaxime resistance	5434	1687	30	13	39.6	0.106
TEM-1	tazobactam resistance	5434	1687	19	8	35.6	0.210
HB36.4	hemagglutinin binding	932	309	127	55	30.6	0.0066
HB80.3	hemagglutinin binding	964	312	83	27	0.51	0.531
Variable for calculations		N	m	n	k		

[a]enrichment $= \left\lfloor \frac{\frac{k}{n} - \frac{m}{N}}{\frac{m}{n}} \right\rfloor \times 100\%$

[b]calculated from a hypergeometric distribution: $P(x=k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$

3.3.6 Strength of the *TEM-1* adaptive mutations for tazobactam and cefotaxime resistance

A comparison with previously reported adaptive mutations for tazobactam resistance indicates that the selective conditions we utilized to identify tazobactam resistance alleles restricts against adaptive mutations with a small effect (74). For the

cefotaxime resistance mutations, we compared our list of adaptive amino acid substitutions with Schenk *et al*.'s (93) extensive list of 48 cefotaxime resistance adaptive point mutations identified from an error prone PCR library of *TEM-1*. Twenty of our 30 amino acid substitutions are unique to our result, and many of these have either not been previously reported or are not previously known to confer resistance in isolation (94). Our selection readily identified all seven adaptive amino acid substitutions previously shown to confer a >3-fold improvement in resistance (93). We find 26 of the 28 possible codon substitutions that can give rise to these seven amino acids substitutions, suggesting our list contains ~93% of all of adaptive alleles with >3-fold improvement in resistance. We identify some substitutions previously shown to confer between 2.6 and 3-fold improvement, and do not identify any substitutions shown to confer <2.6-fold improvement. Substitutions with a ≤3-fold effect represent 85% of the adaptive point mutations (93). In summary, our selection identified a large set of adaptive alleles conferring about 2.5-fold or greater improvement in cefotaxime resistance, which we estimate to be the fittest 15% or more of all adaptive alleles with a single codon substitution.

3.4 DISCUSSION

Provided the three genes we studied are representative of all genes, our results indicate that the standard genetic code possesses a remarkable feature – it provides the advantage of reducing the negative effects of mutations while selectively enriching for adaptive mutations. Thus, although the code's structure limits the exploration of sequence space, it does so in a manner that benefits the evolution of proteins and is a molecular-level

example of how constraints can facilitate evolution (95). We speculate that the architecture of the code results in part from selective pressure for a code that facilitates the evolution of proteins. This evolvability theory on the origin of the genetic code is not mutually exclusive with existing theories and offers additional insight into the origin of the standard genetic code.

Our use of the term "evolvability" and our proposal that the code's enrichment for adaptive mutations provided an adaptive advantage requires further clarification. When compared to other genetic codes that do not enrich for adaptive mutations, the standard genetic code imposes a bias towards adaptive mutations in the standing genetic variation. Of course this bias would not be true for all genes or in all possible environments. Rather, we contend it would be true on average. Thus, provided that adaptation is limited by the supply of adaptive mutations, the standard genetic code would confer a higher degree of evolvability than a code that does not enrich. This advantage the code possesses invokes clade selection, as it provides an advantageous backdrop for adaptive evolution. As evolution proceeds, surviving lineages would become increasingly biased towards those with this code, which experienced more beneficial mutations sooner than their competitors.

Whether this evolvability provided an adaptive advantage (as we propose) or is a byproduct of evolution is a difficult question to answer (96,97). However, our proposal does not suffer from many criticisms of evolvability involving clade selection (97,98). First, we do not need to invoke increased mutation rates or capacity to produce new variation as the source of evolvability. Rather, we contend that the standard genetic code provides a *better* standing genetic variation for evolution than would codes that do not

enrich for adaptive mutations. Thus, early in evolutionary history, this could have provided an adaptive advantage contributing to the standard genetic code winning out over alternative codes that may have been present at the time. Second, the genetic code is not a simple "variability allele" that is prone to being lost by recombination because it is subject to indirect selection. The genetic code is a manifestation of a large set of genes and is central to life. It cannot be lost, and it is difficult to think of ways in which the code would be vulnerable to "selfish" alternatives. Thus, although our results indicate that the standard genetic code possesses the ability to enrich for adaptive mutations today and in the future, we are not invoking a teleological view of evolution. Rather, the adaptive advantage existed earlier in evolution in the context of other competing codes while genomes were smaller and the genetic code exhibited plasticity. We believe that the code's retention of this feature is a testament to how difficult it is to substantially change the genetic code after its fixation in the last universal common ancestor (9).

Our results support the idea that both robustness to error and improved access to adaptive mutations were selected for in the genetic code's evolution. We speculate that there are two possible ways in which our evolvability theory can be reconciled with the adaptive theory vis-à-vis error minimization. (i) First, perhaps a code's error minimization must be balanced by its propensity to promote the evolution of proteins. A code maximized for robustness to error would allow only the most conservative of mutations, which may not be optimal from the perspective of protein evolution. We postulate that a code that allows for the right balance between error minimization and effective exploration of sequence space would be evolutionarily advantageous. In this view the evolvability theory provides a possible explanation for the extent to which, if

any, the code is not optimized for error minimization (6). (ii) Second, it may be that the error minimization and adaptive mutation enrichment provided by the genetic code are two sides of the same coin. Potentially, a conservative genetic code increases the probability of achieving an adaptive mutation by reducing the effect of the mutations (99), consistent with Fisher's Geometric Theorem (100). If error minimization and enrichment for adaptive mutations do come together as a package, an interesting but difficult question to address experimentally is the extent to which each contributed to the origin of the genetic code.

3.5 ACKNOWLEDGEMENTS

We thank Timothy A. Whitehead, Aaron Chevalier, and David Baker for providing their deep sequencing data of the selected and unselected *HB36.4* and *HB80.3* libraries (20). We thank Jeffrey J. Gray and Stephen J. Freeland for constructive comments on the manuscript.

CHAPTER 4

A COMPREHENSIVE, HIGH-RESOLUTION MAP OF A GENE'S FITNESS LANDSCAPE

SUMMARY

Mutations are central to evolution, providing the genetic variation upon which selection acts. A mutation's effect on fitness can be positive, negative, or neutral. Knowledge of the distribution of fitness effects (DFE) of mutations is fundamental for understanding evolutionary dynamics, molecular-level genetic variation, complex genetic disease, the accumulation of deleterious mutations, and the molecular clock. We present a comprehensive DFE for point mutants of the *E. coli TEM-1* β-lactamase gene and missense mutations in the TEM-1 protein. This DFE provides insight into the origin of the genetic code, support for the hypothesis that mRNA stability dictates codon usage at the beginning of genes, an extensive framework for understanding protein mutational tolerance, and evidence that mutational effects on protein thermodynamic stability shape the DFE. Contrary to prevailing expectations, we find that deleterious effects of mutation primarily arise from a decrease in specific protein activity and not protein cellular levels.

4.1 INTRODUCTION

The fitness landscape model for molecular evolution, as first conceptualized by John Maynard Smith in 1970 (1) and generalized by others (101), imagines evolution as a process by which a sequence moves by stochastic processes from its wildtype sequence through fitter and fitter sequences until the sequence reaches a local fitness optimum. The nature of the fitness landscape determines the dynamics of evolution and fundamentally shapes what is and is not possible in evolution. The landscape also defines the relationship between DNA/protein sequence and biological function. Much has been learned from theoretical studies and small-scale interrogations of real fitness landscapes (13). Recent deep mutational scanning studies have provided large datasets relating protein sequence and function (16-20,23-25); however, we still lack a systematic, assumption-free, experimental determination of the DFE for all mutations of a gene performing its native function in its native host. The situation is akin to having a small set of aerial photographs of a geographical area versus having comprehensive satellite coverage such as provided by Google Earth.

We sought to provide a comprehensive, quantitative description of a fitness landscape corresponding to a gene and its nearest neighbors in both DNA and protein sequence space (i.e. the set of all sequences that differ by one base pair, codon, or amino acid) but avoid current limitations of large-scale measurements of fitness. Growth competition experiments or experiments in which alleles are enriched based on a threshold for function are the current state of the art (16-20,23-25). To varying degrees such experiments offer a direct "head-to-head" comparison of alleles but suffer four significant limitations. First, most studies utilize non-native reporter assays (e.g. phage

display, cell surface display, and two-hybrid systems) in which the gene or gene fragment is removed from its native context and host and fused to another gene (but see refs (19,23)). Second, population size can affect the measured value of fitness due to stochastic effects. Third, these experiments have limited ability to measure fitness for low fitness alleles because such alleles are depleted during the course of the experiment. For example, Roscoe et al. (19) were unable to reproducibly measure the fitness of ubiquitin point mutants with a fitness below ~40% of wildtype due to their rapid depletion in growth competition experiments. Thus, although such experiments tell us the location of valleys in the landscape, they cannot tell us anything about what the valleys look like. Fourth, the fitness measurements are subject to the extent and underlying form of genotype-by-environment interactions. For example, the fitness of an antibiotic resistance gene measured by a growth competition experiment will be a function of the arbitrary selective pressure used in the experiment (the antibiotic concentration). Alleles conferring resistance far above or below the level necessary for growth at one antibiotic concentration may show no fitness difference in that environment yet show significant differences at a different antibiotic concentration closer to their resistance limit. We desired to decouple fitness from genotype-by-environment interactions as much as possible to quantify the underlying fitness landscape and thus better understand a gene's intrinsic evolutionary potential and limitations.

4.2 MATERIALS AND METHODS

4.2.1 Description of the band-pass selection system

The band-pass genetic selection for β-lactamase activity can select for *E. coli* cells exhibiting any desired level of β-lactamase activity (42) (Figure 4.1). The two plasmids that comprise the system and the interactions of their components are shown in Figure 4.1A. In this system, cells with too little β-lactamase activity (relative to the amount of β-lactam antibiotic such as ampicillin) cannot grow due to the β-lactam's inhibitory effect on cell wall synthesis. However, the β-lactam also serves to confer resistance to the antibiotic tetracycline (Tet) via induction of the *ampC* promoter. Cells with too much β-lactamase activity (relative to the amount of β-lactam antibiotic) cannot grow in the presence of Tet since rapid degradation of the β-lactam prevents sufficient induction of TetC expression. Thus cells challenged to grow in the presence of Tet and Amp will grow only if the amount of β-lactamase activity is balanced between having enough activity to degrade the β-lactam to allow cell wall synthesis but not too much to prevent induction of TetC (Figure 4.1B). Since increased β-lactam concentration shifts the amount of β-lactamase activity necessary for growth to higher levels, our cells can be used to select for any level of β-lactamase activity (from the absence of activity to full activity) simply by adding the necessary concentration of the β-lactam. A more in depth characterization of this system is provided in Sokha et al. (42) including the linear relationship between a cell's minimum inhibitory concentration (MIC) of Amp in the absence of Tet and the concentration of Amp in the presence of Tet that provides the best growth for that cell. This linear relationship holds over three orders of magnitude of Amp concentration. In the experiments reported here, fitness corresponds to the Amp

concentration providing the best growth in the presence of Tet, which is roughly 25% of the MIC_{Amp}.

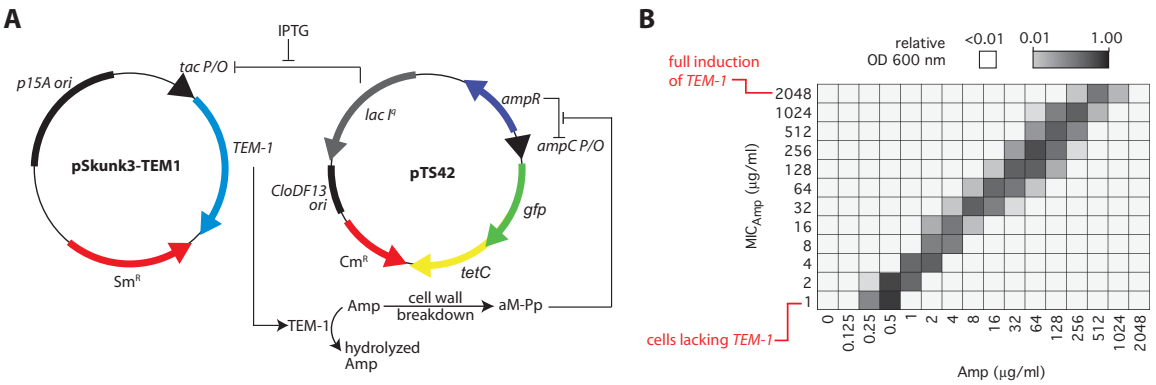


Figure 4.1. Bacterial band-pass filter for β-lactamase activity. (A) Essential components of the genetic circuit for band-pass selection (42). In the absence of sufficient cellular TEM-1 β-lactamase activity for hydrolysis of ampicillin (Amp), cell wall synthesis is compromised and cells cannot proliferate. Cell wall breakdown results in the accumulation of aM-pentapeptide (aM-Pp), which induces the *ampC* promoter via interactions with AmpR (102,103) resulting in the production of TetC (which confers tetracycline resistance) and the green fluorescent protein (GFP). The level of Amp necessary to induce *ampC* is lower than the level that prevents the growth of *E. coli* cells (103). Thus, cells that hydrolyze Amp too efficiently cannot grow in the presence of Tet. As a result, in the presence of tetracycline and Amp, cells will proliferate only if they possess an intermediate amount of Amp hydrolysis activity. The level of Amp hydrolysis activity necessary for growth increases linearly with Amp concentration (103). *TEM-1* expression is regulated through IPTG-induction of the LacI-repressed *tac* promoter. (B) Demonstration of band-pass selection for BLA activity in *E. coli* SNO301 cells (42). The growth of cells in liquid LB media containing 20 µg/ml Tet was detected by measuring the OD at 600 nm of the liquid culture and is presented in the form of a heat map as a function of Amp concentration (x-axis) for cells with different levels of β-lactamase activity (y-axis, as measured by the minimum inhibitory concentration (MIC) for Amp).

4.2.2 Fitness determination

The method for measuring fitness makes use of the band-pass genetic selection system (Figure 4.1) and is schematically depicted in Figure 4.2. Comprehensive codon mutagenesis library CCM2 comprises three separately constructed libraries, one for each

third of the gene (74). Collectively, they are designed to contain all possible single codon substitutions in the *TEM-1* gene. Each library was plated on LB-agar plates containing 50 μg/ml spectinomycin, 50 μg/ml chloramphenicol, 300 μM IPTG, 20 μg/ml tetracycline and 13 different ampicillin concentrations (2-fold increments of ampicillin ranging from 0.25 μg/ml to 1024 μg/ml) at a cell density of 1700 CFU/cm^2. Plates were incubated at 37°C for 20 hrs. Colonies were recovered from the 39 plates with LB broth, and plasmid DNA isolated using the Qiagen QIAprep Spin Miniprep kit (27106). The plasmid DNA was linearized by restriction endonuclease digestion with SphI and purified using the Zymo DNA Clean & Concentrator kit. Because the plasmid miniprep also contained the band-pass plasmid, pTS42, the concentration of the library plasmid, pSkunk3-CCM2, was determined by running a sample of the linearized DNA on an ethidium bromide agarose gel and analyzing the band intensities of the two respective linearized plasmids. The mass ratio of pTS42:pSkunk3-CCM2 was determined to be ~12.5.

PCR amplicons of each of the 39 sub-libraries were created using Titanium Lib-A "A" and "B" fusion primers that included a 10-base MID barcode identifying the sub-library from which the DNA originated. In order to minimize the rate of recombination or "PCR-jumping" between DNA template strands in the PCR amplification, experiments were performed to find the minimum amount of template DNA and minimum number of PCR cycles necessary to obtain sufficient PCR product. Each 25 μl PCR reaction had 22.4 pg (5x10^6 molecules pSkunk3-CCM2) linearized template DNA, 0.5 μM each barcoded primer, 200 μM each dNTP, 1X HF Phusion buffer, and 0.5 units Phusion high-fidelity DNA polymerase. Cycler conditions were 98°C for 30 sec, 25 cycles of 98°C for 10 sec, 61.9°C for 15 sec, 72°C for 3 min, and then 72°C for 5 min. PCR product DNA

concentration for each reaction was determined using the Quant-iT Picogreen dsDNA Assay kit (P7589). The barcoded amplicons were then mixed together in molar proportion to the number of colonies that grew on their respective sub-library selection plate. A total of 2.4 μg of the amplicon mixture was electrophoresed on a TAE 0.7% agarose gel and then gel purified using the QIAquick Gel Extraction Kit (28706). Further purification was performed using the Agencourt AMPure XP PCR Purification kit (A63880) to remove short DNA fragments, primers, and primer dimers. The final purified amplicon DNA concentration was determined using picogreen and diluted to 10^9 molecules/μl in 1X TE and then further diluted to 10^7 molecules/μl in DI water. 454 sequencing was performed by Tufts University Core Facility on a Roche 454 GS FLX+ instrument.

We created custom MATLAB scripts to analyze the raw 454 sequencing reads. From a combination of five full or partial plate runs, we obtained a total of 1,325,979 reads. We aligned the reads to the template sequence, sorted the reads by barcode, and mapped all codon substitution mutations, ignoring indels (which are a common sequencing error). Reads were filtered out if the average base call quality score was less than 30, the read did not span the entire mutagenesis region, or if there was more than one codon mutation per read. We obtained 772,296 reads that passed our filtering requirements and had a single codon substitution.

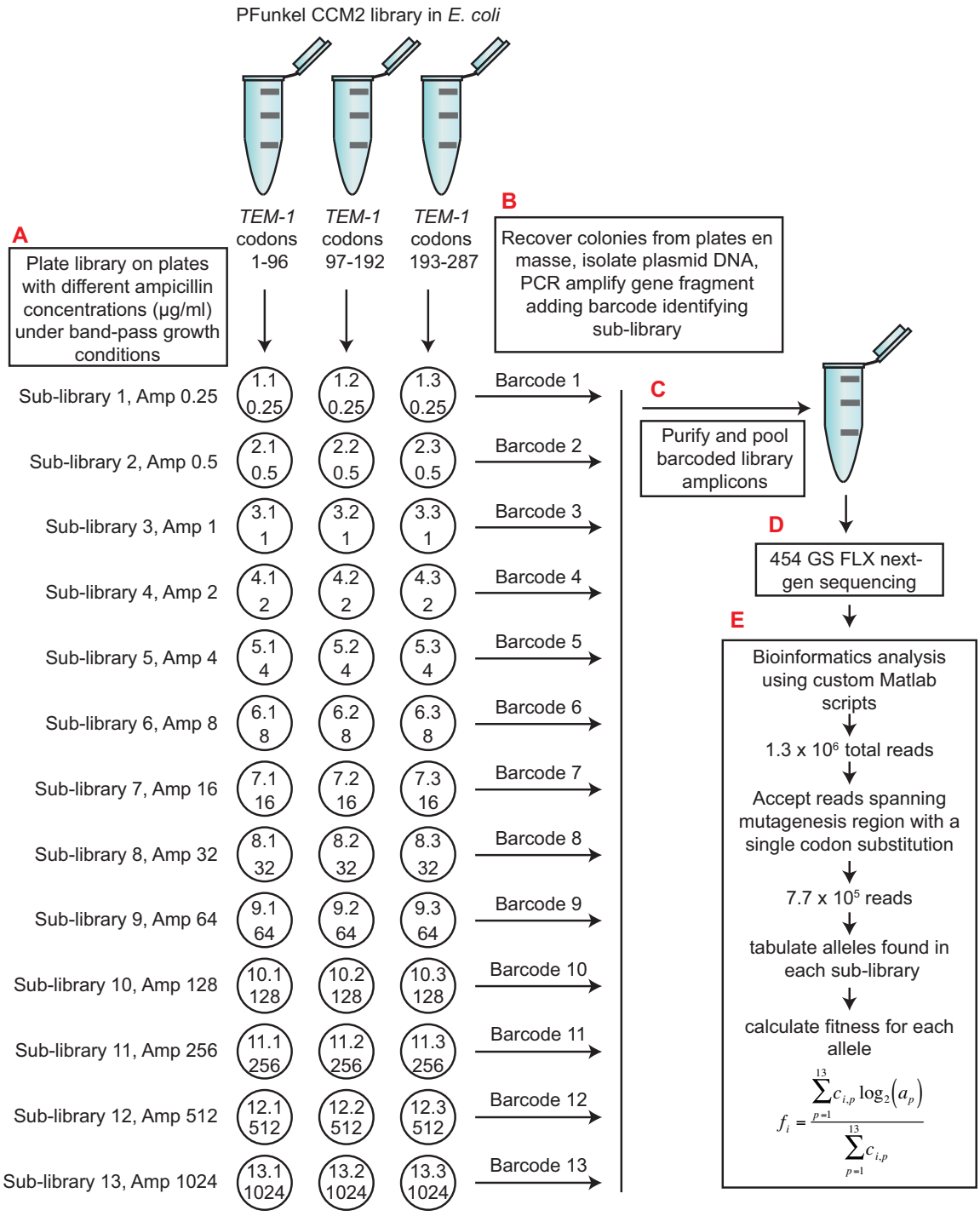


Figure 4.2. System for measuring fitness of *TEM-1* alleles. The CCM2 library in strain SNO301 (separate libraries for each third of the *TEM-1* gene) was plated on media containing Tet and different concentrations of Amp. PCR products with barcodes identifying the plating conditions were subjected to 454 GS FLX DNA sequencing and the counts of alleles on each of the growth conditions was used to quantify fitness.

We tabulated the number of sequencing counts for each allele in each sub-library (see online supporting information Data S1). Since the distribution of growth as a function of Amp is roughly symmetric when plotted as the $\log_2$(Amp concentration) (42), we determined the unnormalized fitness f of allele i as

$$f_i = \frac{\sum_{p=1}^{13} c_{i,p} \log_2\left(a_p\right)}{\sum_{p=1}^{13} c_{i,p}} \qquad \text{Equation 1}$$

in which $c_{i,p}$ is the number of counts of allele i on sub-library plate p in the deep sequencing data and a_p is the concentration of Amp on sub-library plate p in μg/ml . Counts of a particular allele can be expected to appear on 3-4 adjacent sub-library plates (42), corresponding to an 8-16 fold window of Amp concentration. In a minority of cases however, counts were observed in more than four sub-libraries or counts were observed to cluster in non-adjacent sub-libraries. Two phenomena can account for this: sequencing errors and the presence of an unintended, fitness-altering mutation outside the sequencing region. The frequency of sequencing errors is low relative to that of mutations, since 87% of the library members contain a mutation, and sequencing errors should occur with equal frequency among all sub-libraries. We implemented several measures to facilitate appropriate fitness assignment. First, an allele was assigned a fitness value only if its sequence was observed at least five times. Second, the window of the four adjacent sub-libraries with the highest combined sequencing counts was identified and only these values were included in the fitness calculation. Third, if an allele presented with multiple clusters of counts, the cluster corresponding to the fitness closest to the average fitness of the other synonymous codons was selected. For alleles with mutations in the start codon,

knowledge about alternative *E. coli* start codons was used to assign fitness. For alleles with nonsense mutations with multiple clusters of counts, the lowest cluster was used to assign fitness. For the fitness of missense mutations, we combined the sequencing counts of the corresponding synonymous codons (see online supporting information Data S2) and recalculated the fitness using the same methods as above for codon substitutions.

We normalized all fitnesses by the fitness of wildtype as follows:

$$w_i = \frac{2^{f_i}}{2^{f_{WT}}} \qquad \text{Equation 2}$$

This result is a normalized fitness w_i that is 1.0 for wildtype *TEM-1*, > 1.0 for beneficial mutations and between 0 and 1.0 for deleterious mutations. We determined the fitness of wildtype *TEM-1* (f_{WT}) using Equation 1 using the counts of all alleles with a synonymous substitution in *TEM-1*, since the fitness of these varied very little. As a check, we compared this value to the fitness determined by Equation 1 using the counts of all sequencing reads that lacked a mutation. The two values differed by only 2.5%. Fitness values are tabulated in online supporting information Data S1 and Data S2.

We determined an upper limit on the error in our fitness measurements. We assumed (solely for the purpose of this error determination) that synonymous mutations have no fitness effect. We compared an allele's fitness to the mean of all alleles with a synonymous mutation at the same position and expressed this difference as a percentage of the mean. The distribution of values for this 'percent difference in fitness' did not vary with fitness, indicating our fitness measurements are equally accurate at low and high fitness values (Figure 4.3A,B), unlike in growth competition experiments. As expected, the width of the distribution narrowed with the number of times the allele was observed

in the deep sequencing results (the 'allele count') (Figure 4.3C,D). We used the standard deviation of this distribution as a function of allele count as an estimate of the error in fitness. This error is an upper limit since synonymous mutations can have fitness effects (104).

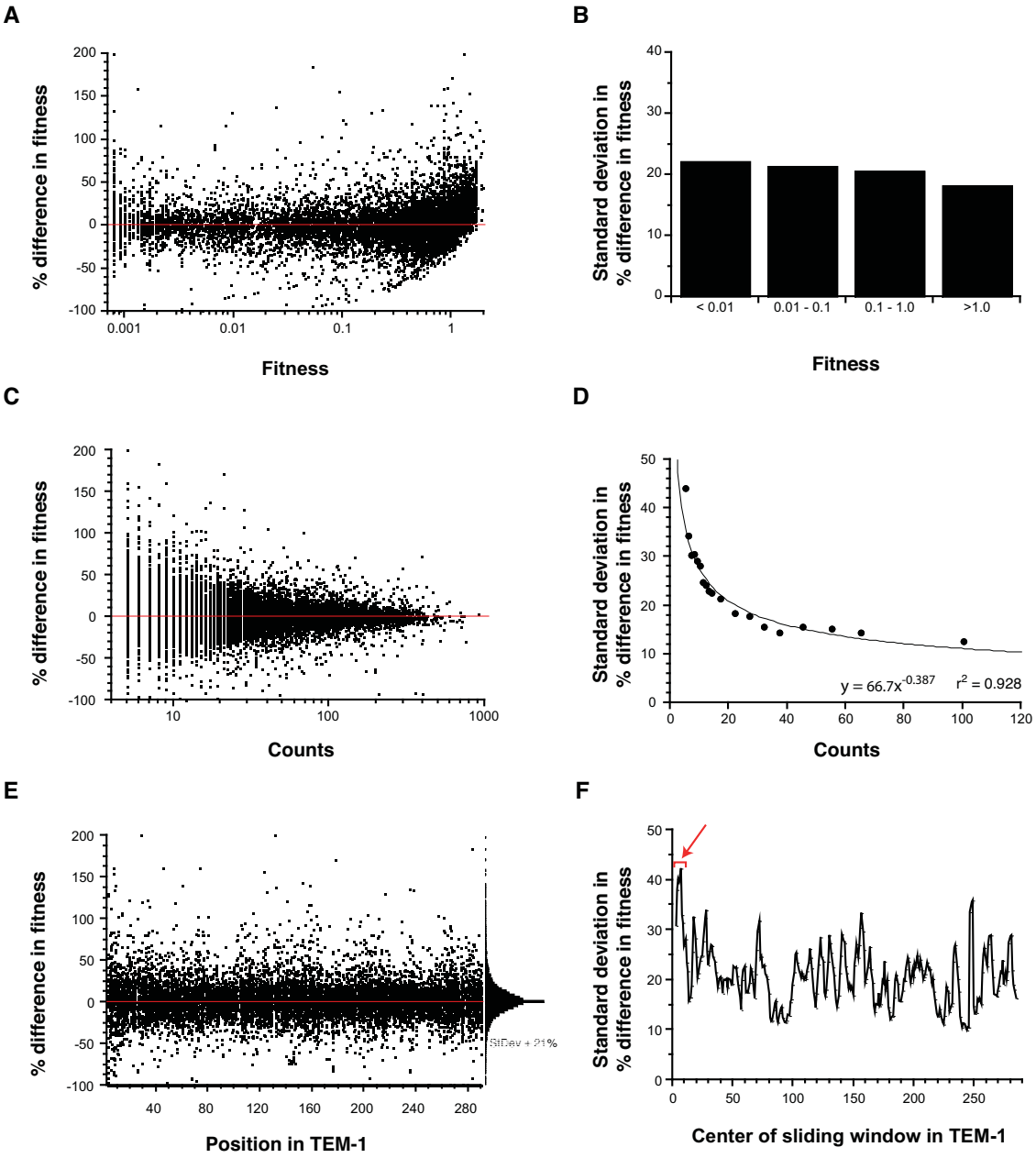


Figure 4.3. Distribution of synonymous effects. We determined the percent difference in fitness between the fitness of allele *i* with a mutation at codon *j* of the gene and the

mean fitness of all alleles with a synonymous mutation at codon *j* (including allele *i*). (A) The distribution of percent fitness difference as a function of fitness indicates that the fitness measurement is equally precise at low and high fitness values. (B) This observation is further illustrated by the standard deviation of percent fitness differences for different fitness ranges. (C) The percent fitness difference decreases with the number of times an allele is encountered in the deep sequencing experiment (i.e. the counts). (D) The standard deviation in percent fitness difference as a function of counts, defined here, was used as an upper limit of error in the fitness measurements. (E) The percent fitness difference is fairly uniform across the *TEM-1* sequence. (F) However, a broader distribution in the first ten codons of the gene is apparent. A sliding window of three positions was used.

4.2.3 Prediction of mRNA stability at the transcript start

The RNAfold utility of the Vienna RNA Package (version 2.1.2) was used to predict the minimum free energy of RNA sequences (105). For each allele in Figure 4.10, the Gibbs free energy was calculated as the average free energy of every 39 nt window centered on nucleotides from -5 to +10 of the gene start as described (106).

4.2.4 Mutational tolerance

The observed effective number of amino acids k_o^* at a position was determined from the fitnesses of the *n* missense mutations with fitness data at that position using Equations 3 and 4.

$$S = -\sum_{i=1}^{n} w_i \log_2 w_i \qquad \text{Equation 3}$$

$$k_o^* = 2^S \qquad \text{Equation 4}$$

We obtained the effective number of amino acids k^* by normalizing k_o^* to be based on 20 amino acids by Equation 5.

$$k^* = \frac{20k_0^*}{n} \qquad \text{Equation 5}$$

A table of k_o^* and k^* is provided in the online supporting information Data S4.

4.2.5 **Prediction of protein thermodynamic stability**

PyRosetta v3.4.0 r55307 (107) was used to compute the difference in score (in Rosetta Energy Units, REU) between the mature structures (lacking the signal sequence) of each amino acid mutant and wild type TEM-1 (Protein Data Bank identifier 1XPB (108))). The score is designed to capture the change in thermodynamic stability caused by the mutation ($\Delta\Delta G$) (109). First, all side chains were repacked (sampling from the 2010 Dunbrack rotamer library (110)) and minimized for the wild type structure using the talaris2013 scoring function. Next, each missense mutation was introduced and all residues within a 10 Å distance of the mutated residue's center were repacked followed by a linear minimization of the backbone and all side chains. This procedure was performed 50 times, and the predicted $\Delta\Delta G$ is the average of the three lowest scoring structures. The PyRosetta script used has been included as part of the standard PyRosetta installation (http://www.pyrosetta.org/dow) and can be found in the apps directory of the PyRosetta installation as delta_score_per_mutation.py. PopMusic predictions of $\Delta\Delta G$ (Figure 4.17B) were determine online at http://babylone.ulb.ac.be/popmusic (111).

4.2.6 **Preparation of samples for protein dose and total catalytic activity assays**

For analysis of the sub-libraries, the three libraries of CCM2 were combined and plated on 13 different Amp concentrations as in the fitness measurement experiment described above. Individual clones were plated on the same Amp concentration upon which they were found (i.e. 8 or 16 μg/ml) at the same colony density at which the original library grew (i.e. the numbers of the colonies on the plates were the same). Colonies were recovered by sweeping with LB broth containing 15 v/v% glycerol and 2

w/v% glucose. Aliquots comprising cultures with equal cell density were pelleted and the supernatant removed. After freezing at -80 ^{o}C for 10 min, the cells were lysed in 250 µL BugBuster Protein Extraction Reagent (Novagen 70584-3) containing ≥0.25 U/µL benzonase nuclease (Sigma E1014-25KU), 3 units/µL rLysozyme (Novagen 71110), and 1% protease inhibitor cocktail (Sigma P8849). Samples were incubated with gentle shaking at 4^{o}C for 30 min, then centrifuged at 14,000 rpm for 30 min at 4°C. The supernatant was recovered as the soluble protein fraction and the pellet resuspended in 250 µL 8M urea as the insoluble fraction. Samples were aliquoted and stored at -80°C. Total protein concentration of each lysate sample was measured using the DC protein assay (Bio-Rad 500-0111) with a BSA standard.

4.2.7 Protein dose quantification

Western blots for each sub-library were performed with 10 µg total protein of each lysate sample. A standard curve was prepared by diluting the lysate from sub-library 13 in 2-fold increments, with increasing additions of control lysate from cells not expressing any *TEM-1* allele to maintain a constant amount of 10 µg total protein. SDS-PAGE gels (Novex NP0323BOX) were electrophoresed for 45 min at 190 V, and transferred to a PVDF membrane (Bio-Rad #162-0177) for 30 min at 15 V. The membrane was blocked with a 4% milk solution in PBST (1X PBS, 0.05% tween 20) for 1 hr at room temperature with shaking. The primary anti-TEM-1 mouse monoclonal antibody (Thermo MA1-20370, 500-fold dilution in blocking solution) was incubated for 2 hrs at room temperature or overnight at 4^{o}C with shaking. The membrane was then washed three times with PBST for 5 min with shaking. The secondary goat anti-mouse antibody (Bio-Rad 170-5047, 20,000-fold dilution in blocking buffer) was incubated for

1 hr with shaking at RT. The membrane was then washed three times with PBST for 5 min with shaking, and a final wash with 1X PBS. A total of 1 ml of chemiluminescence detection reagent (Bio-Rad #170-5070) was then applied to the membrane, incubated for 1 min, and then imaged on a Bio-Rad Gel Doc XR system, recording exposures at 5 sec intervals. Representative westerns are shown in Figure 4.18. Quantity One 1-D analysis software (Bio-Rad) was used to quantify the band intensity for BLA sub-library or clone samples. For each western blot, the image of longest exposure before any band reached detector saturation was selected for quantification. The local adjusted volume parameter was used as the measure of band intensity because it subtracts the local background around each band. Intensity was then converted to relative protein concentration using the standard curve correlation with the same corresponding exposure time.

4.2.8 Measurement of total catalytic activity

Catalytic activity of the sub-libraries and clones was determined by measuring nitrocefin hydrolysis rates. A solution of 50 μM nitrocefin in 10 mM phosphate buffer pH 7.4 was incubated in a 96-well plate (BD Falcon 353072) at 37^{o}C for 2 min. Then 1.0 to 2.5 μL of the soluble fraction of the lysate was added to a final volume of 200 μL and mixed by pipetting up and down briefly. The rate of hydrolysis was measured as the initial slope of absorbance as a function of time as measured at 486 nm at 37°C in a SpectraMax Plus 384 Microplate reader. The initial rate was normalized by the total amount of protein added for each sample.

4.2.9 Theoretical calculation of total catalytic activity vs. fitness

Allele fitness results from the ability to hydrolyze Amp at the concentration of Amp present on the plate. Our experimental measure of total catalytic activity uses the

initial rate of nitrocefin hydrolysis at 50 μM nitrocefin. These will not correlate precisely 1:1 owing to the differences in catalytic constants for hydrolysis of the two substrates and the fact that nitrocefin hydrolysis is measured at a set nitrocefin concentration but fitness is evaluated at different Amp concentrations. However, we can predict the form of the relationship. We assumed Michaelis-Menten kinetics and the following values for k_{cat} and K_m for Amp (1187 s^{-1} and 42.7 μM) and nitrocefin (917 s^{-1} and 128 μM), which are average values for TEM-1 of that measured in several previous studies. We first calculated the reduction of total Amp hydrolysis activity expected for the expected mean fitness of the sub-library. We then assumed that this reduction in catalytic activity comes from equal percent reductions in the k_{cat} and K_m. We then assumed that this same percent reduction in the catalytic constants will occur for nitrocefin hydrolysis. We then calculated the predicted relative initial rate of nitrocefin hydrolysis at 50 μM nitrocefin using the Michaelis-Menten equation. We performed these calculations (a) assuming the protein dose was the same for all fitness values and (b) using the experimentally observed measurements of protein dose shown in Figure 4.16B. The results are presented as Figure 4.4.

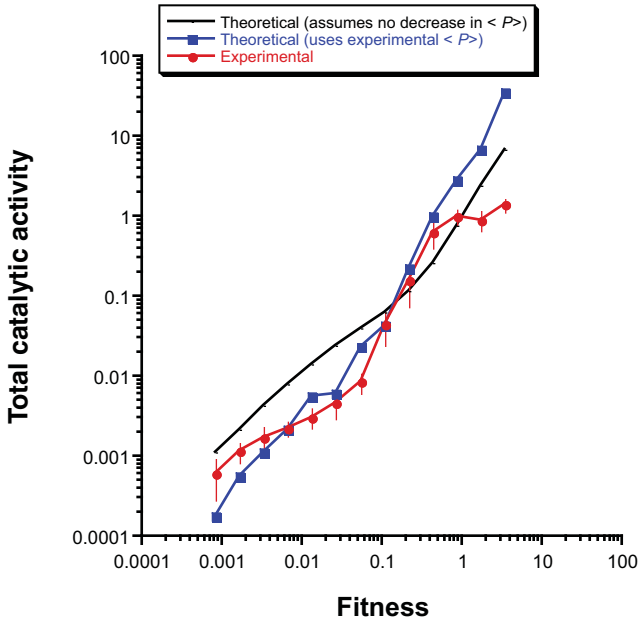


Figure 4.4. Expected relationship between fitness for Amp resistance and total cellular catalytic activity as measured by nitrocefin hydrolysis. The slight sigmoidal relationship is illustrated by the theoretical calculation that assumes that protein dose does not change with fitness (black line). By instead using the actual measures of protein dose (blue squares), the inflection point shifts to lower values and the theoretical curve more closely matches the experimental data (red circles).

4.3 RESULTS AND DISCUSSION

4.3.1 The fitness landscape of *TEM-1* β-lactamase

We chose to measure the DFE of the *E. coli TEM-1* β-lactamase gene, a convenient model for the study of evolution and the fitness effects of mutations (112). *TEM-1* confers high resistance to penicillin antibiotics such as ampicillin (Amp). Thus, when *E. coli* cells bearing *TEM-1* are challenged to grow in the presence of Amp, alleles conferring an enhanced ability to degrade the antibiotic will enrich. Thus, Amp resistance, as typically measured by the minimum inhibitory concentration (MIC) of the antibiotic, is an excellent proxy for fitness (75,76,113), although the approach may not

completely capture minor fitness differences not associated with antibiotic resistance. However, MIC assays suffer the drawbacks of being low-throughput and low-resolution. Alleles with known mutations must be isolated and tested individually, and MICs are measured in discrete values (typically 2-fold increments). These limitations are exemplified by a recent study of the effect on amoxicillin resistance of 18% of the possible amino acid substitutions in TEM-1 in which the resolution of MIC values was insufficient to capture the effects of synonymous mutations or to identify any beneficial mutations (113). Here we describe a synthetic biology approach to quantify fitness of *TEM-1* in a single experiment that avoids or ameliorates the limitations of growth competition experiments and MIC assays. Additionally, since *TEM-1* increases the Amp resistance of *E. coli* cells over 1000-fold, the combination of *TEM-1* and Amp afforded the opportunity to determine the DFE over a wide range of fitness values (113).

We determined the DFE for 98.2% (2536/2583) of all point mutations (i.e. all 1-bp changes) and 83.9% (15,167/18,081) of all codon substitutions in the *TEM-1* gene (Figure 4.5). The latter includes all 1-, 2-, and 3-bp changes of the 287 codons of *TEM-1*. We also determined the DFE for 95.6% (5212/5453) of the possible single amino acid substitutions in the corresponding TEM-1 protein (Figure 4.6). The source of *TEM-1* variants was a previously described library (CCM-2) designed to contain all possible single codon substitutions in the *TEM-1* gene (i.e. each codon position in the gene could be changed to any of the other 63 codons but each allele had only one position changed) (74). To measure fitness, we first partitioned the CCM-2 library into 13 partially overlapping sub-libraries based on relative Amp resistance using a synthetic gene circuit that functions as a tunable band-pass genetic selection for Amp resistance (42) (Figure

4.1). Next, we performed deep sequencing on each of the sub-libraries, counting how many times each allele appeared in each sub-library and used these statistics to quantify each allele's conferred antibiotic resistance or fitness (*w*) relative to *TEM-1* (Figure 4.2). We used the fitness effects of synonymous mutations to determine an upper limit on the error of our fitness measurements (see section 4.2.2 and Figure 4.3). Our method enables the accurate fitness quantification of any allele and avoids population size effects because the alleles are isolated on plates and the probability of observing an allele in the experiment is not a function of its conferred fitness. Additionally, the method decouples fitness from genotype-by-environment interactions, at least as far as the major environmental factor affecting fitness is concerned (i.e. the antibiotic concentration). We sequenced 27 randomly selected alleles from two of the sub-libraries and found these alleles' fitness values were in the expected range (Figure 4.19).

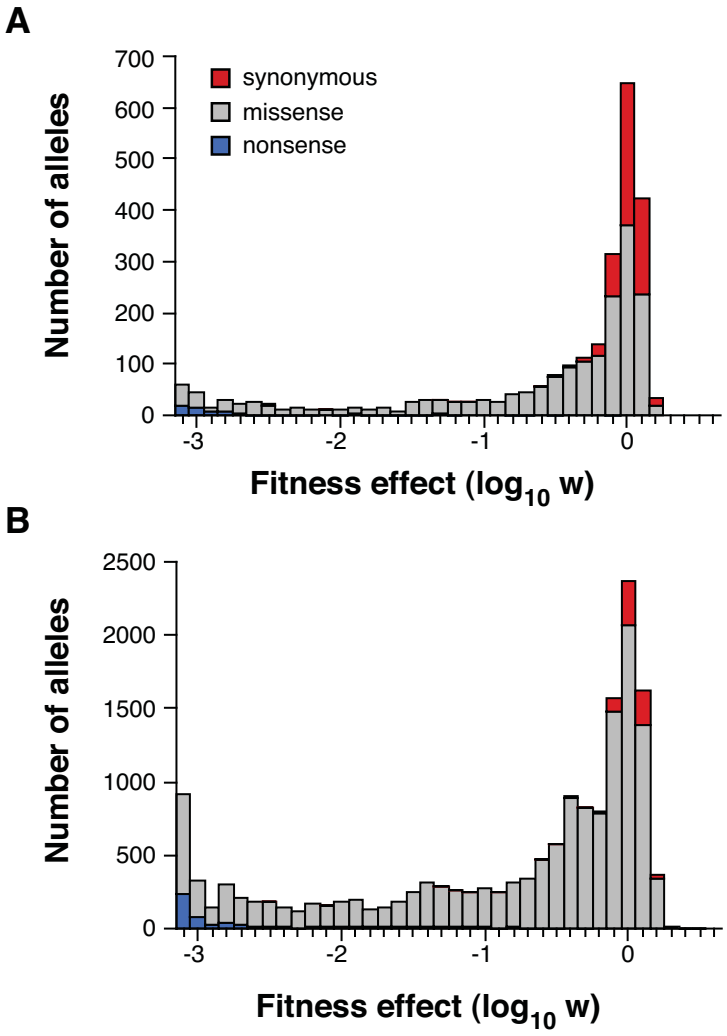


Figure 4.5. Distribution of fitness effects (DFE) of mutations in *TEM-1*. (A) The DFE of point mutations (i.e. 1-bp changes in the gene). (B) The DFE of all possible codon substitutions (i.e. all 1-, 2- and 3- base changes in the 287 codons of *TEM-1*). Fitness values for conferring ampicillin resistance are presented on a log scale with 0 corresponding to the fitness of *TEM-1*. The contributions of synonymous (red), missense (grey), and nonsense (blue) mutations to the DFE are indicated. Fitness as a function of codon substitution is provided in the online supporting information.

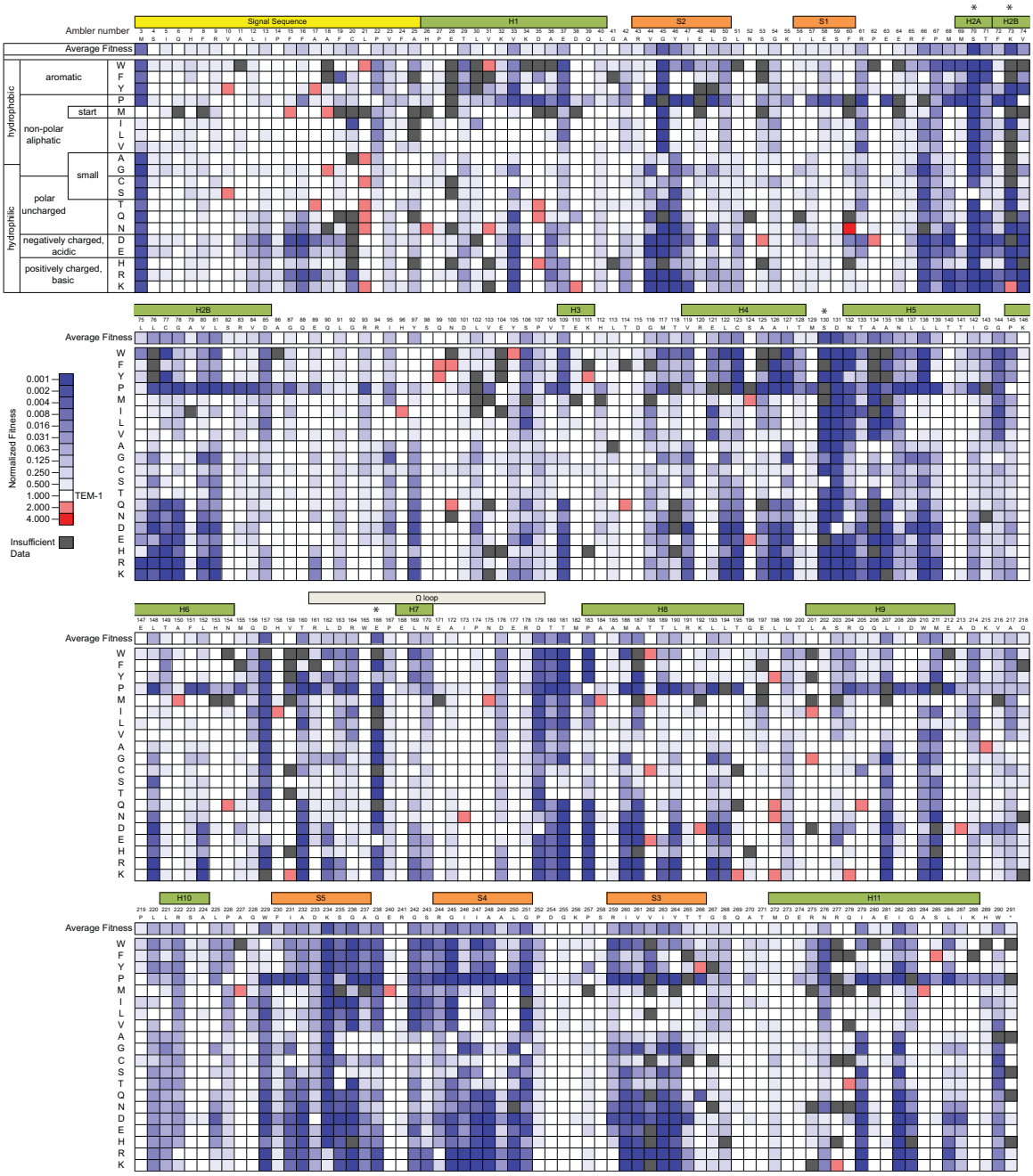


Figure 4.6. The sequence-function landscape of TEM-1. The heat map indicates the fitness values for ampicillin resistance of the indicated amino acid substitution. The Ambler consensus numbering system (61) for class A β-lactamases is used. An asterisk indicates key active site residues. For the start codon, fitness values correspond to the average of the codons for the indicated amino acid though methionine is expected to be the amino acid incorporated. Fitness as a function of missense mutation is provided online in supporting information.

TEM-1 was fairly robust to mutations (Figure 4.5). Nearly half (47.3%) of all alleles retained at least 50% of the fitness of *TEM-1*. Among alleles with point mutations, 63.8% maintained at least 50% of the fitness of *TEM-1* (53.2% of the non-synonymous and 97.2% of the synonymous point mutations). Still, a sizable fraction of the mutants lost >90% of their fitness (19.6% of the point mutations and 30.3% of all codon substitutions), roughly in line with previous estimates of the frequency of mutations having severe deleterious effect (114). Among point mutants, only 6% of the alleles completely lost the ability to provide any Amp resistance and 33% of those were nonsense mutations (Figure 4.5A). Only 7.1% (1,074/15,167) of the alleles and 7.0% (367/5,212) of the missense mutations increased fitness above that of *TEM-1* outside the range of the error. The bimodal distribution was qualitatively similar to the DFE of randomly chosen point mutations in DNA and RNA viruses (15,115), the DFE of induced mutations in yeast (22), the DFE of missense mutations of ubiquitin (19), samplings of *TEM-1*'s DFE for amoxicillin resistance (113), and estimations of TEM-1's DFE for Amp resistance (112).

4.3.2 The benefits of the genetic code's architecture

TEM-1's DFE supports two theories on the origin of the genetic code. The first theory, the adaptive theory, states that the genetic code is arranged to minimize the deleterious effects of mutations and mistranslations (3,7). This theory predicts that point mutations would be less deleterious than 2- or 3-bp substitutions. We have recently shown this prediction held true for mutations in two small genes (*HB36* and *HB80* (20)) that were reengineered for a new function in a non-native organism (116). Here, we find that this prediction is also true of a wildtype gene in its native host. The median changes

in relative fitness for 1-bp, 2-bp, and 3-bp substitutions at a codon position were -0.36, -0.52, -0.63, respectively. More significantly, the frequency of point mutations among the alleles with a fitness less than 0.1 was 35.3% less than that expected if the point mutations were evenly distributed across all fitness values (P = 1.1 x 10^{-40} based on comparison to a hypergeometric distribution). Error minimization was more pronounced with *HB36* and *HB80*; point mutations were 56.4% and 53.8% depleted from clones with a fitness less than 0.1, respectively (P = 1.35 x 10^{-18} and 1.27 x 10^{-21}) (116). We postulate that *TEM-1* is an "evolutionarily mature" gene that is more robust to mutations including non-conservative mutations, whereas the re-engineered genes can be thought of as genes undergoing evolution to a new function and are thus less robust to mutation. This result suggests that the code's error minimization might have higher importance early in a gene's evolution to a new function.

TEM-1's DFE also supports our recently proposed evolvability theory, which states that the standard genetic code's architecture enriches for adaptive mutations and that this enrichment provided an adaptive advantage over alternate codes early in evolutionary history (116). Among the 367 beneficial missense mutations in TEM-1, 41.1% can be achieved by point mutations, 32.5% higher than the 31.0% expected if 367 missense mutations were chosen at random (P = 8.8 x 10^{-6} based on comparison to a hypergeometric distribution). Our comprehensive analysis of adaptive mutations in a natural gene in its native host is the strongest evidence yet that the code's arrangement makes adaptive mutations more likely. However, the role such enrichment played in the origin of the genetic code and whether or not the enrichment is a side effect of the code's error minimization bias are difficult questions to answer (116).

4.3.3 The effects of synonymous mutations

The effects of synonymous mutations on protein synthesis and fitness have important implications for evolution and biotechnology. However, despite an abundance of plausible hypotheses, we lack a mechanistic understanding of these effects (104). Our systematic strategy provides an assumption-free approach for testing and generating these hypotheses. We first examined the fitness of 725 alleles synonymous to *TEM-1*. Beneficial and deleterious synonymous mutations distributed differently across the sequence of *TEM-1* (Figure 4.7A). Beneficial mutations occur primarily in positions 15-30 and 130-260, whereas deleterious mutations appeared in clusters in the first half of the gene and were almost absent from the second half of the gene. No trend in the types of substitutions for either beneficial or deleterious effect was apparent other than eight of the ten beneficial mutations at Arg codons being to the rare *E. coli* codons AGA (2/10) and AGG (6/10). The pattern of beneficial and deleterious synonymous codons indicates the existence of regions of *TEM-1* with sub-optimal and less robust mRNA properties, respectively.

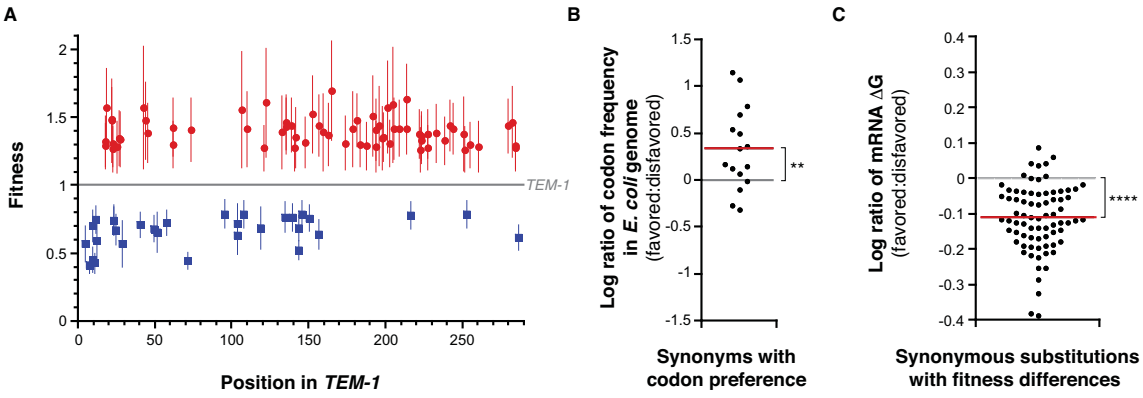


Figure 4.7. Effects of synonymous substitutions. (A) Beneficial and deleterious synonymous mutations in *TEM-1* are not evenly distributed. Alleles synonymous to *TEM-1* with a fitness significantly higher (red circles) or lower (blue squares) than that of *TEM-1* are shown. The criteria for significance was that the error did not extend into the range fitness = 1 ± 0.1. Error bars provide an upper estimate on error on the fitness measurements as described in the text. (B) An analysis of pairs of synonymous alleles with mutations in codon positions 2-10 of the gene (Figure 4.8 and 4.10) revealed that codon preferences tended to be for codons with a higher frequency in the *E. coli* genome. (C) Preferred codons at positions 2-10 of the gene were predicted to result in mRNA with less stable structures around the initiation codon. Red bar is the mean. ** $P < 0.01$, **** $P < 0.0001$ by Student's *t*-test.

We next analyzed the effects of 14,055 synonymous substitutions among the set of 15,167 alleles with fitness measurements (Figure 4.3). Over the length of the entire gene, CUA (Leu), AGG (Arg), and UCG (Ser) provided an average fitness advantage over some of their synonymous codons (Figure 4.9), but the advantage was only ~5%. Interestingly, CUA and AGG are rare codons in *E. coli*. Codon usage often differs in the beginning of the gene from the rest of the gene, which has been hypothesized to result from a selection against 5' mRNA structure and/or a selection for rare codons that provide a slower elongation time at the 5' end (104). We addressed both these hypotheses with our data. Positions 2-10 in *TEM-1* had an almost 2-fold broader distribution of synonymous effects compared to any other section of the gene (Figure

4.3F). Within these nine positions, we observed 26-85% mean fitness increases for certain codons of Ala, Arg, Gly, Leu, Pro, and Ser relative to select synonyms (Figure 4.8). These synonymous fitness differences distributed differently among the nine positions (Figure 4.10). Contrary to the slow elongation hypothesis, favored codons tended to appear more frequently in the *E. coli* genome than their corresponding disfavored codon (Figure 4.7B) (117). However, none of the 16 observed codon preferences were between the most and least frequently used codons within a synonym set suggesting that codon usage was an inadequate explanation for the observed preferences (e.g. as a result of tRNA abundance). We next calculated the folding energy of the mRNA around the initiation codon for alleles exhibiting fitness differences (105). In almost all cases, favored codons reduced mRNA stability around the translation start site compared to disfavored codons (Figure 4.7C). Our finding supports a recent study that concluded that mRNA structure at the beginning of genes, and not codon usage, determines the translation rate (106).

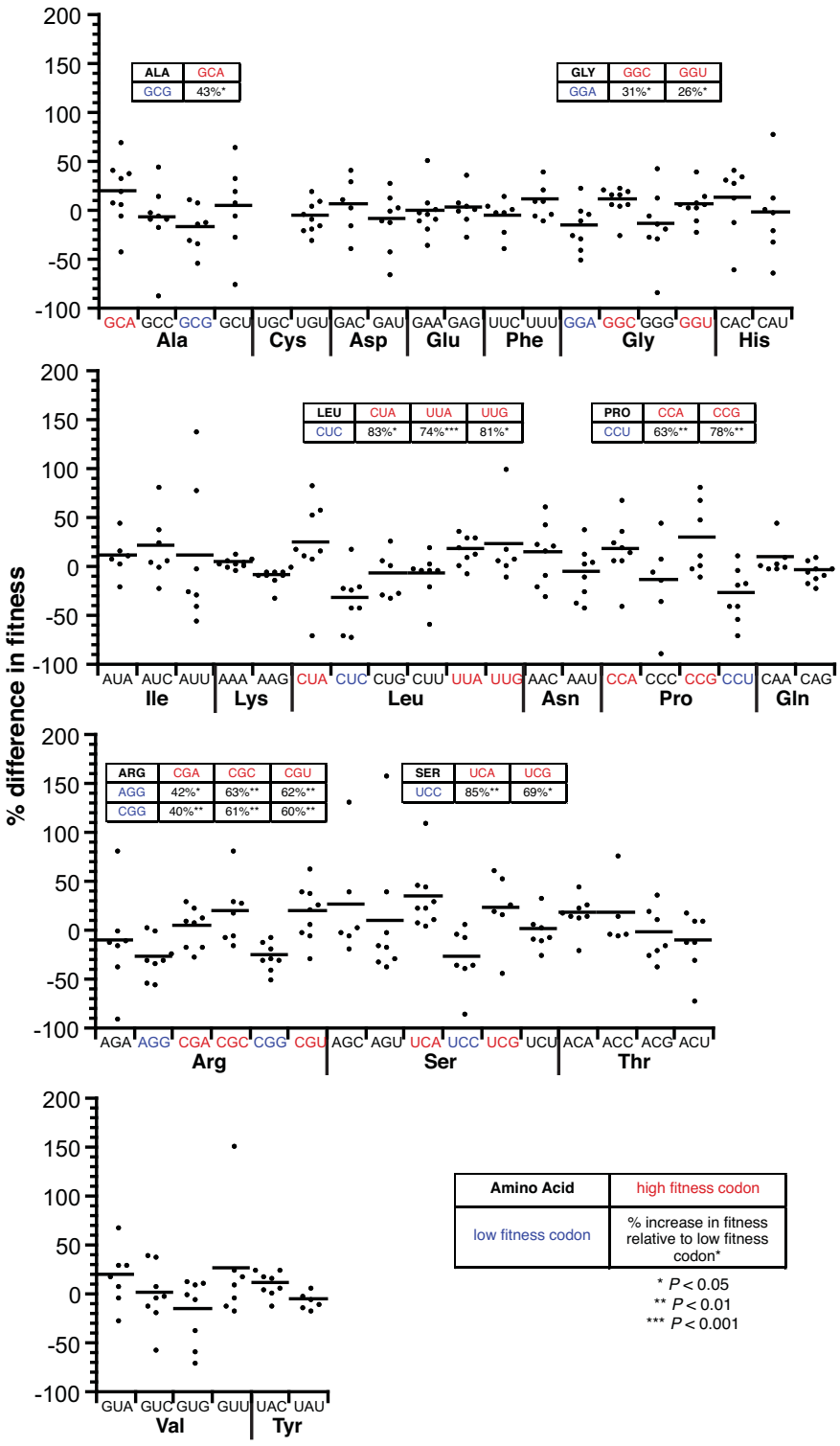


Figure 4.8. Fitness effects of codon usage at positions 2-10 in *TEM-1*. The percent fitness differences of synonymous substitutions within positions 2-10 in *TEM-1* were analyzed as a function of the codon substituted. *P* values were determined by Student's *t*-test.

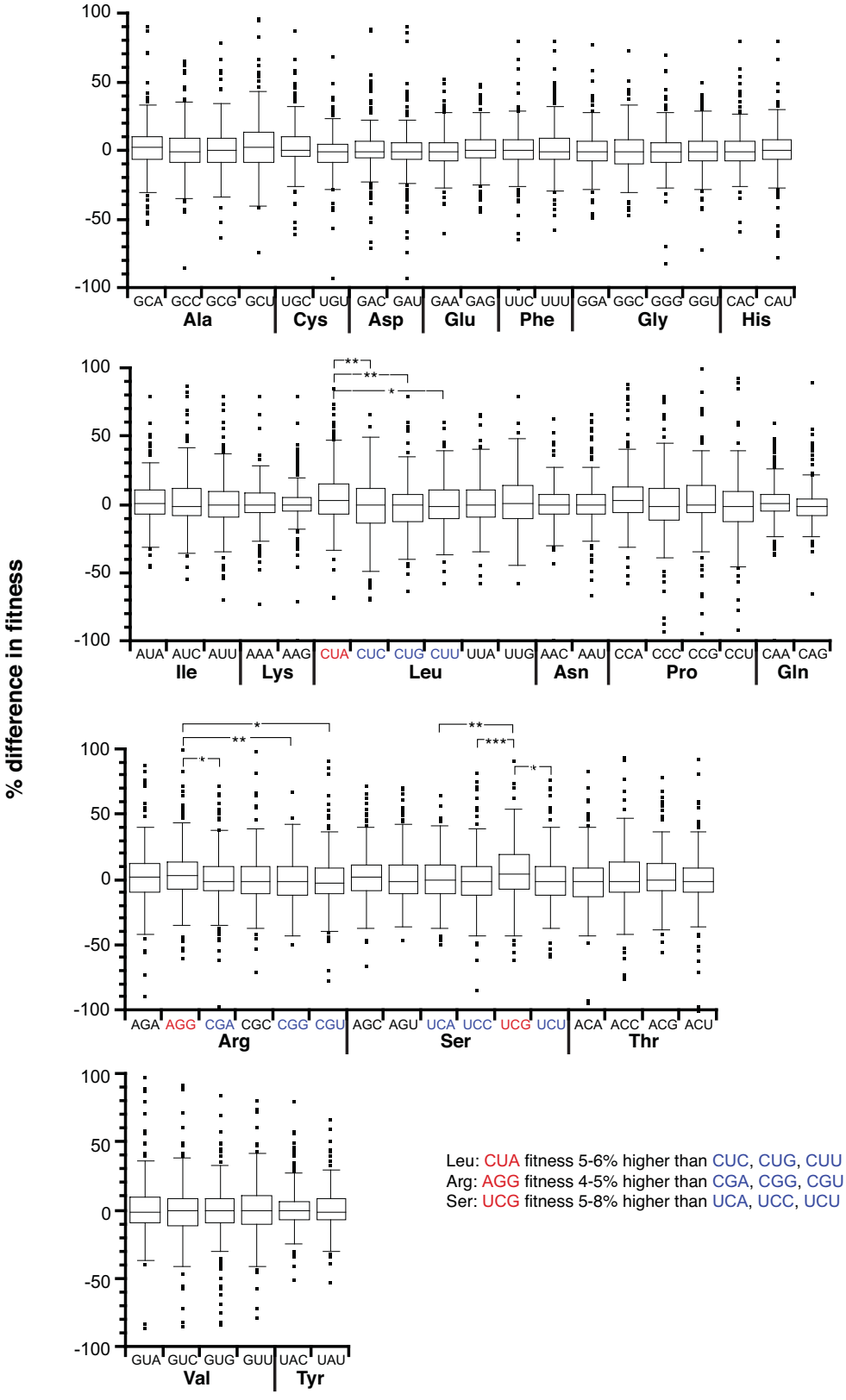


Figure 4.9. (legend on next page)

Figure 4.9. Global fitness effects of codon usage in *TEM-1*. The percent fitness differences of 14,055 synonymous substitutions among the 15,167 alleles with fitness measurements were analyzed as a function of the codon substituted. The global bias along the entire *TEM-1* gene for any particular codon, if there is any, is on the order of 5% or less. Our analysis for global codon bias is more likely to be able to identify smaller differences in mean fitness for codon sets with a greater number of codons. Thus, although we identify about 5% fitness differences between certain codons for Leu, Arg, and Ser, the fact that these are the three amino acids with six codons made the identification of significant but small differences in these codon sets more likely. Equal differences may or may not exist within other codon sets but we lack the statistical significance to identify them. *P* values were determined by Student's *t*-test.

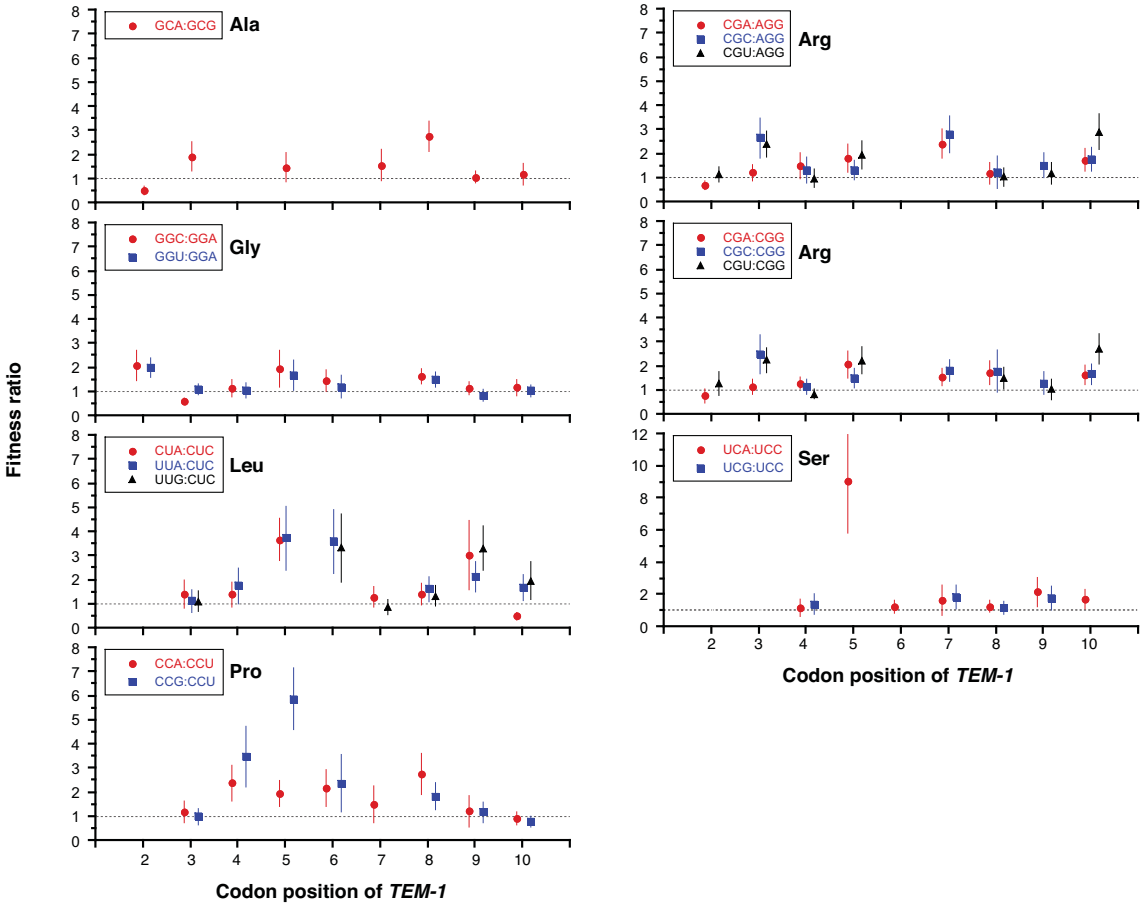


Figure 4.10. Positional dependence of synonymous fitness effects at positions 2-10 in *TEM-1*. For the synonymous codon pairs determined to have significant differences in fitness within positions 2-10 of the gene (Figure 4.8), the ratio of fitnesses for the two alleles is shown as a function of position in the gene.

4.3.4 Observed exceptions to the standard genetic code

Among the three stop codons, UAG (amber) exhibited nonsense suppression (Figure 4.11A). A 3' flanking purine after the UAG enhanced this suppression (Figure 4.11B), as has been observed with the amber suppressor tRNA allele *supE* (118,119). We sequenced the seven tRNAs known to serve as amber nonsense suppressors and found that *E. coli* strain SNO301 harbors the *supE44* allele, which consists of a duplicate copy of the *glnV* tRNA gene, *glnX*, with the expected anti-codon mutation (thereby inserting glutamine at UAG codons) (120). This allele suppressed a UAG with a 3' flanking purine at a mean efficiency of 7-10% (Figure 4.11C). Substitutions for the AUG start codon that provided significant fitness (>5% of that of *TEM-1*) included seven of the nine point mutants of AUG (Figure 4.12), consistent with known native and non-native alternative initiation codons in *E. coli* (121,122). In addition we observed that GUA, GUC, and GUU could serve as weak initiation codons (7-14% as efficient AUG). Initiation from GUA in *E. coli* has been previously reported (123), but initiation from GUC and GUU has not.

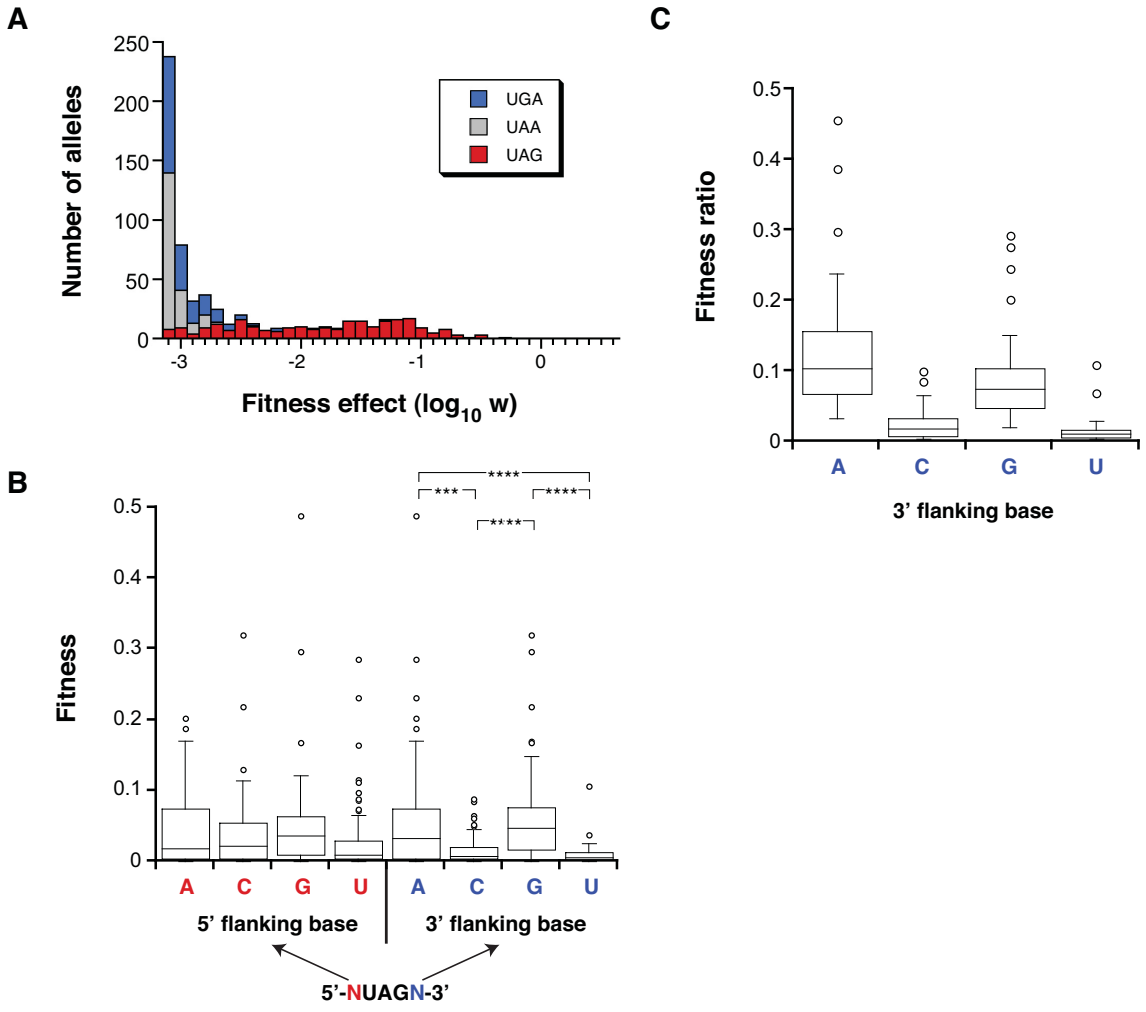


Figure 4.11. Fitness effects of nonsense mutations in *TEM-1*. (A) The DFE of nonsense mutations as a function of the three nonsense codons. Fitness values are presented on a log scale with 0 corresponding to the fitness of *TEM-1*. (B) The efficiency of nonsense suppression at UAG is higher if an A or a G is at the 3' flanking position. *** $P = 0.0002$, **** $P < 0.0001$ by Wilcoxon–Mann–Whitney test. (C) The efficiency of nonsense suppression at UAG is strongly determined by the 3' flanking nucleotide. The fitness ratio compares the fitness of an allele with a mutation to UAG to the fitness of an allele with a missense mutation to glutamine at the same position. Only glutamine missense mutations with $w > 0.25$ were considered. The median efficiencies were 10.4% (3' A) 1.8% (3' C), 7.5% (3' G) and 1.1% (3' U). Differences between A/G and C/U have a P value of <0.0001 (Wilcoxon–Mann–Whitney test).

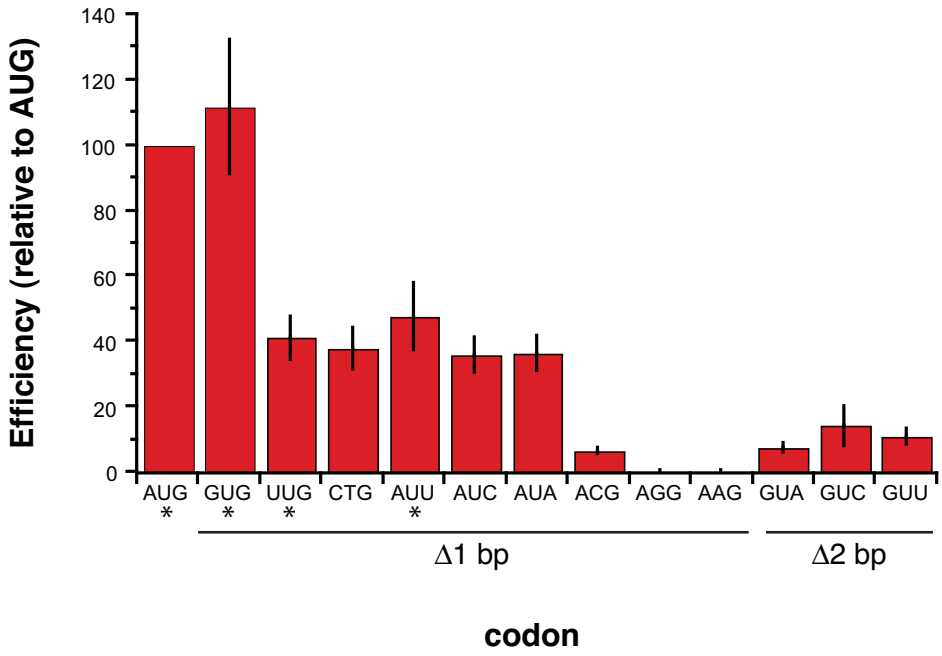


Figure 4.12. Relative efficiency at which select codons serve as initiation codons. The efficiency was determined by dividing the fitness of the allele with the indicated codon at position 1 in the gene by the fitness of *TEM-1* (i.e. with AUG in position 1). Asterisks indicate known native initiation codons in the *E. coli* genome. All codons that differ from AUG by 1 bp are shown as are the three codons that differ by more than 1 bp that exhibited >1% efficiency.

4.3.5 Mutational tolerance and the effects of missense mutations

Since the effects of nonsynonymous mutations dwarfed that of synonymous mutations, we combined the fitness data of synonymous codons to determine the DFE of missense mutations in the TEM-1 protein (Figure 4.6). The fitness landscape of TEM-1 broadly matched what is known about protein structure in general and TEM-1 in particular. For example, proline was the least tolerated substitution (see Figure 4.13, which displays TEM-1's amino acid substitution matrix for Amp resistance), especially in alpha helices, and key TEM-1 active site residues did not tolerate mutation (Figure 4.6). *TEM-1*'s signal sequence is required for export via the Sec pathway to the periplasm. The signal sequence (Figure 4.6) tolerated most mutations consistent with the pathway's

broad specificity (124). However, the hydrophobic core of the signal sequence did not tolerate substitution of charged residues, consistent with typical export-defective mutants in Sec pathway signal sequences (124). Signal sequence residue L21 was a hot spot for beneficial mutations, and L21F is found in some extended-spectrum resistant TEM alleles (125).

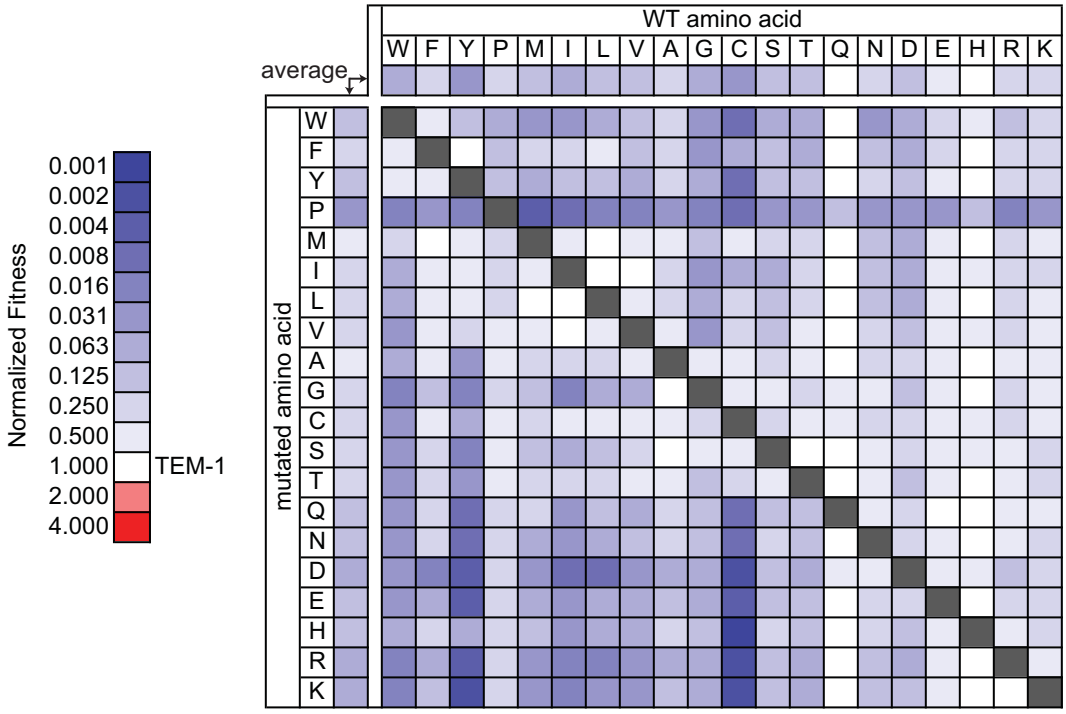


Figure 4.13. Amino acid substitution matrix for TEM-1. The heat map indicates the average fitness for the indicated substitution. Tabulated data for this heat map is provided in online supporting information Data S3.

The comprehensive fitness landscape of missense mutations enables a rigorous determination of a protein's mutational tolerance in its biological context. We determined the effective number of amino acids at a position (k^*), which derives from the fitness entropy that is calculated from the distribution of fitness values for the 20 amino acids at that position. This measure of tolerance is more informative than establishing an arbitrary fitness cutoff for deciding whether or not a mutation is tolerated. Our approach

is analogous to how information-theoretical entropy is used to measure variability at a position in a set of aligned sequences (126). A k^* value of 1 corresponds to a position at which all missense mutations completely inactivate the protein and a k^* value of 20 means that all 19 amino acid substitutions provide the same fitness as the wildtype amino acid. The distribution of k* was strongly biased towards high values (Figure 4.14A). Half of all positions accepted 15.5 or more amino acid substitutions. Under the simple assumption of a linear correlation, percent solvent-accessible surface area accounted for 49% of the k*'s variance (Figure 4.14B) and predicted k^* better than distance from the active site (Figure 4.14C) or a k^* determined from a sequence alignment of 156 class A β-lactamases (23) (Figure 4.14D). Both a k^* based on the sequence alignment (Figure 4.14D) and previous calculations of k^* for TEM-1 (23) (Figure 4.15) greatly underestimated TEM-1's mutational tolerance. The latter's underestimation presumably derives from that study's use of alleles with multiple mutations without accounting for epistatic effects between the mutations and the high stringency used in selecting functional sequences. The tolerance of amino acid position i weakly correlated with positions i+1, i+3, and i+4 (correlation coefficient 0.25-0.28, $P \leq 1.8 \times 10^{-5}$) but not i+2 or i > 4. This correlation primarily occurred at residues with high k^* values. The eight positions with $k^* < 2.5$ include the four strictly conserved residues involved in the catalytic mechanism (S70, K73, S130 and E166) and four other highly conserved residues (Figure 4.14E). In contrast, the 42 most tolerant positions ($k^* > 19$) predominantly appeared away from the active site in surface loops and at position 2 in alpha helices (Figure 4.14E). Alpha helices (mean $k^* = 13.5 \pm 5.4$) tolerated substitutions

better than beta strands (mean $k^* = 9.89 \pm 4.8$) ($P = 0.0005$ Student's t-test), perhaps a reflection of the buried nature of the beta-strands.

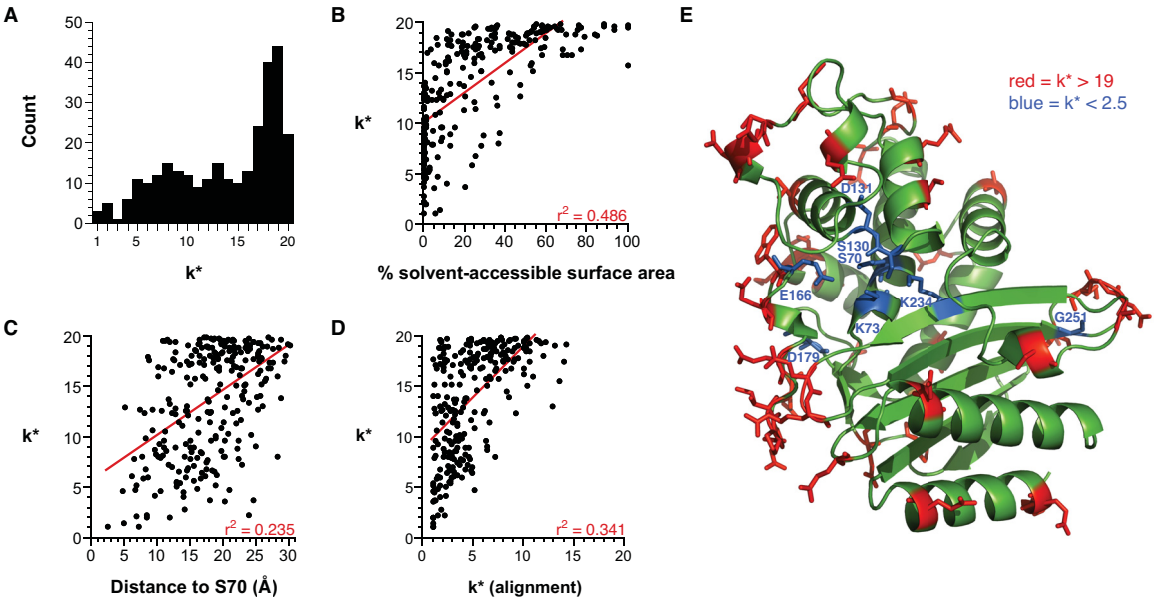


Figure 4.14. Tolerance of TEM-1 to missense mutation. Tolerance was measured by the effective number of amino acids at a position (k^*). (A) The distribution of k^* values in TEM-1. (B) Correlation of k^* correlates with percent solvent-accessible surface. (C) Correlation of k^* with distance from the active site. (D) Correlation of k^* with a sequence alignment of 156 class A β-lactamases (23). (E) Model of TEM-1 (PDB ID 1XPB (108)) indicating the least tolerant positions (k*<2.5, shown in blue), which include the key active site residues S70, K73, S130 and E166, and the most tolerant positions (k*>19, shown in red).

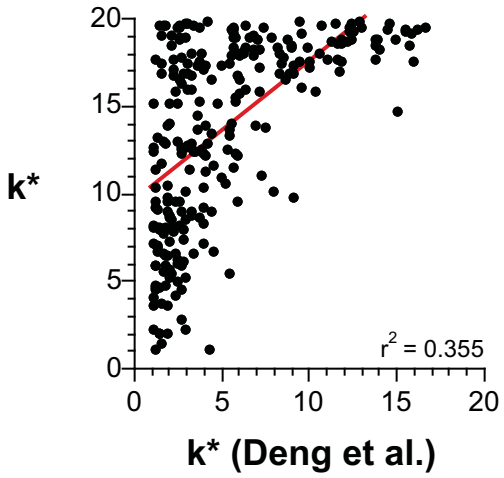


Figure 4.15. Comparison of k* of this study with that determined from *TEM-1* alleles with multiple mutations by Deng et al. (23).

4.3.6 The determinants of mutational effects on fitness

What basic phenomena underlie the DFE? TEM-1 offers a simple case for addressing this question, as TEM-1 fulfills a single cellular role (inactivation of β-lactam antibiotics), and the reaction's substrate and product are not part of any native *E. coli* metabolic or signaling pathway. For *TEM-1*, the dominant determinant of fitness is the total antibiotic hydrolysis activity in the cell, v_t, which is a product of the TEM-1 protein's specific catalytic activity (the rate at which it hydrolyzes the antibiotic, v_{sp}) and how much protein is present in the cell in a correctly folded, soluble form (the protein dose, P (112)):

$$w \propto v_t = v_{sp} P \qquad \text{Equation 6}$$

Experimental evolution studies have shown that v_{sp} and P are equally important targets for adaptive evolution (127,128). Protein dose is expected to be a function of the thermodynamic stability (ΔG) as well as protein production rates (i.e. arising from mRNA properties) and degradation rates (i.e. proteolytic susceptibility). Both computational and experimental studies show that, on average, missense mutations decrease thermodynamic stability (129). A prevailing hypothesis on the origin of deleterious fitness effects of mutation states that thermodynamic stability is the primary determinant of the DFE through its effect on protein dose (114,129-131). Although the hypothesis is intuitive and appealing, experimental evidence for a significant correlation between protein stability and fitness via an effect on protein dose is scant. Mutations that reduce function often show decreased protein dose (132,133); however, mutations that increase stability can reduce specific activity (134) and reductions in protein stability often accompany adaptive mutations (135). A small-scale study of 25 point mutants of

TEM-1 showed no correlation between predicted $\Delta\Delta G$ and conferred Amp resistance (76); however, a study of 990 point mutants of TEM-1 found that 15-19% of the variance in amoxicillin resistance could be explained by the computationally predicted change in protein stability caused by the introduction of the mutation in TEM-116 (TEM-116 is TEM-1 with the V84I and A184V mutations) (113). A comprehensive, systematic study of (a) the relationship between fitness and ΔG for a native protein, and (b) the relative contributions of protein dose and specific activity to the deleterious effects of mutations would more fully address the fundamental phenomena underlying the DFE.

We predicted $\Delta\Delta G$ ($\Delta G_{wildtype} - \Delta G_{mutant}$) using Rosetta (107,109) for 4783 missense mutations of TEM-1, allowing limited backbone flexibility (Figure 4.16A). Variants that were predicted to be more stable tended to have higher fitness (Figures 4.16A and 4.17A). The larger a mutation's deleterious effect on fitness, the higher the probability that the mutation produced a very large predicted energy score (Figure 4.17A). Predictions of $\Delta\Delta G$ using PopMusic (111), a more empirical approach to predicting changes in protein stability than Rosetta, produced similar results (Figure 4.17B). Fitness also correlated with mutational effects on melting temperature (Figure 4.17D). Our results provide the strongest experimental evidence yet that effects on protein stability significantly shape the DFE.

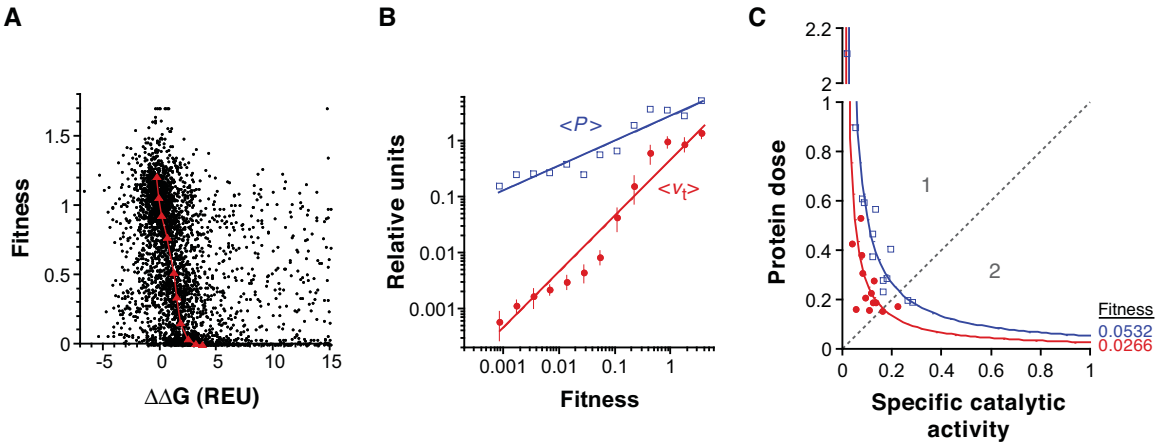


Figure 4.16. The determinants of fitness. (A) Loss of fitness correlates with loss of thermodynamic stability. Fitness is shown as a function of change in ΔG as predicted by Rosetta (107) for 4783 missense mutations of TEM-1. The median predicted ΔΔG for fitness deciles is shown in red triangles. Predicted changes >15 Rosetta energy units (REU) are not shown and are not considered in the median calculation. The distribution of ΔΔG for select fitness deciles can be found in Figure 4.17. (B) Total cellular catalytic activity determines *TEM-1* fitness. The average total cellular catalytic activity $\langle v_t \rangle$ and the average protein dose $\langle P \rangle$ were experimentally measured for 13 sub-libraries of ~15,000 unique *TEM-1* alleles partitioned based on relative fitness. The values of $\langle v_t \rangle$ and $\langle P \rangle$ are relative to that of *TEM-1*. The slight sigmoidal form of $\langle v_t \rangle$ is an expected artifact of the methodology (Figure 4.4). The error bars represent the standard deviation of six assays from two independent experiments. The lines are guides for the eye. (C) Fitness phase space defined by Equation 6. The protein dose and specific catalytic activity (relative to *TEM-1*) of 26 randomly selected members of sub-libraries 6 (red solid circle) and 7 (blue open square) is shown. The dotted line corresponds to an equal decrease in protein dose and specific catalytic activity. In region 1, a mutation affects specific catalytic activity more than protein dose. The solid lines are of constant fitness at the average expected fitness values of the two sub-libraries from which the alleles were randomly selected.

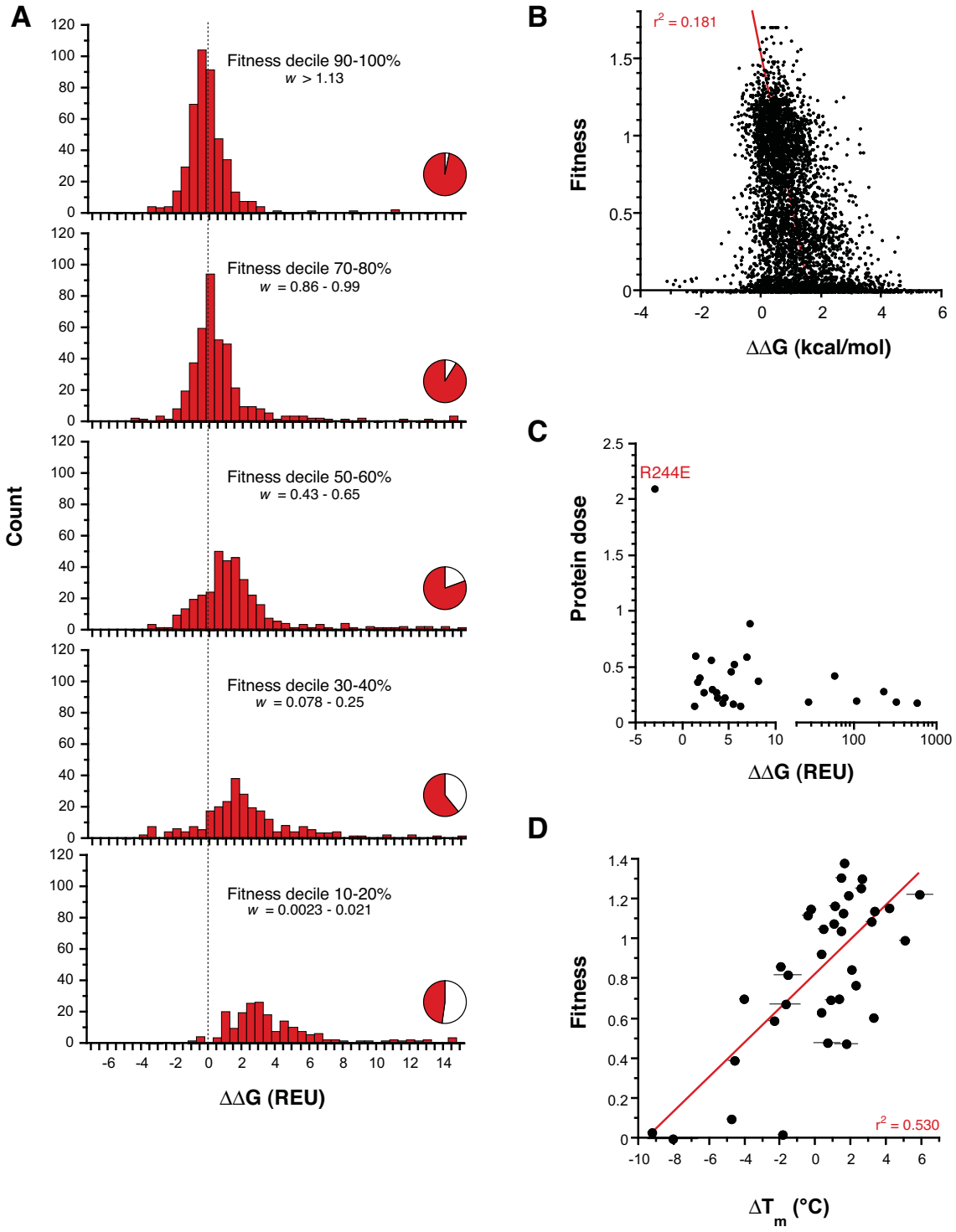


Figure 4.17. (legend on next page)

Figure 4.17. The correlation between fitness and protein stability. (A) Distribution of predicted ΔΔG (by Rosetta) for select fitness deciles of 4783 missense mutations of TEM-1 (i.e. the data of Figure 4.16A). In general, proteins with reduced fitness are predicted to have decreased stability. The fitness decile (i.e. 90-100% indicates the fittest 10%) and the decile's corresponding fitness range are indicated. The pie graphs indicate in red the fraction of ΔΔG values for a decile that are <15 Rosetta energy units (REU). Limitations in the accurate prediction of ΔΔG (136) including the constraints on backbone movement contribute to the high ΔΔG values of some variants. (B) Fitness is shown as a function of change in ΔG as predicted by PopMusic (111) for 4783 missense mutations of TEM-1. (C) Protein dose as a function of predicted ΔΔG (by Rosetta) for 27 randomly selected alleles with low fitness (i.e. the alleles of Figure 4.19). Protein dose is expressed relative to TEM-1. All alleles were predicted to lose thermodynamic stability, the exception being R244E, which had a 2.1 higher protein dose than TEM-1 that is likely the result of the increase stability. (D) Fitness as a function of experimentally measured change in the melting temperature (T_m). We compared the fitness and melting temperature (T_m) of 36 TEM-1 mutants (23,135,137-142). The least fit sequences tended to be those few sequences that had lost more than 4 °C in T_m. Unfortunately, the dataset is biased towards stabilizing mutations obtained by functional selection that generally had a <2-fold effect on fitness.

Whether or not mutational reductions in protein dose are the major cause of loss of fitness has not been experimentally addressed. We experimentally addressed this question by analyzing the soluble fraction of cell lysates of sub-libraries and randomly selected alleles from our *TEM-1* library. We first established that w and v_t are directly proportional as predicted by Equation 6 by measuring the mean total hydrolysis activity of the cell $\langle v_t \rangle$ for the 13 sub-libraries of CCM-2 (Figure 4.16B). We measured protein dose by quantitative western blot of the soluble fraction of cell lysates (Figure 4.18). We assumed all soluble protein was folded and active. The mean protein dose $\langle P \rangle$ of the sub-libraries did not decrease nearly as rapidly with decreasing fitness as $\langle v_t \rangle$ did (Figure 4.16B). In addition, an increase in the mean amount of aggregated protein did not accompany a loss of fitness (Figure 4.20). These findings suggest that mutational effects on v_{sp} rather than on P may play the larger role in the deleterious effects of mutation. Since this interpretation hinges on the distribution of values of P in the sub-libraries, we

measured P for 27 randomly selected alleles with a fitness of about 0.025-0.05. We chose this fitness range so that the mutational effects were substantial, but not inactivating. This ensured that our conclusions would not depend on small changes in w and P. All 27 alleles exhibited a decrease in both protein dose and predicted thermodynamic stability relative to TEM-1 with the exception of the R244E allele, which showed an increase in both (Figure 4.17C). From w and P we calculated v_{sp} using Equation 6 and examined the fitness phase space by plotting P versus v_{sp} (Figure 4.16C). We find that the deleterious effects of mutations arise more from a decrease in specific activity than from a decrease in protein dose. Despite the large negative effects on specific activity, the mutated residues of the 27 alleles were not clustered around the active site but were scattered throughout the protein (Figure 4.19C). Thus, the dominant effect of mutation on specific activity does not arise because the 27 mutations were biased to be proximal to the active site. Thus, we postulate that mutational effects on specific activity are as important to the DFE at high fitness as at low fitness, but this postulate requires experimental investigation.

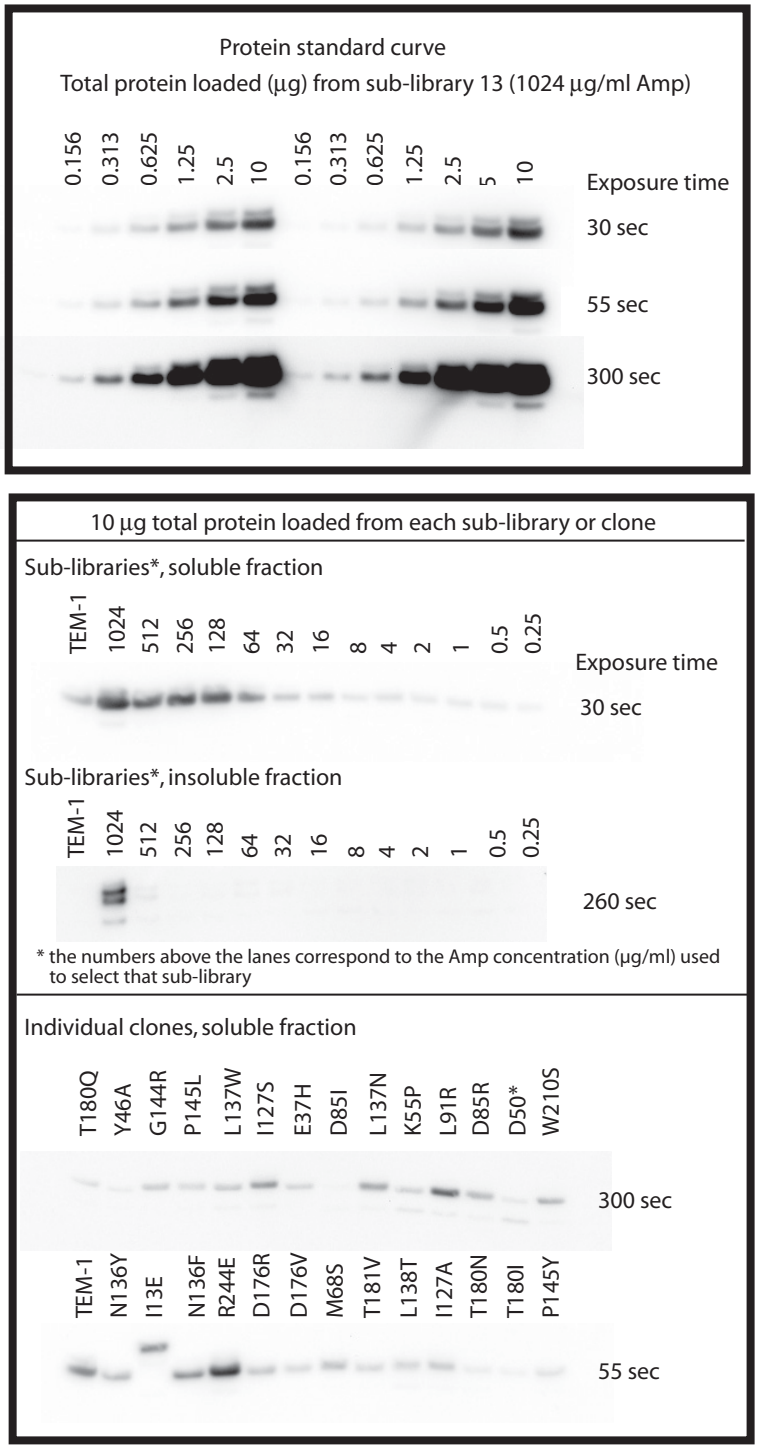
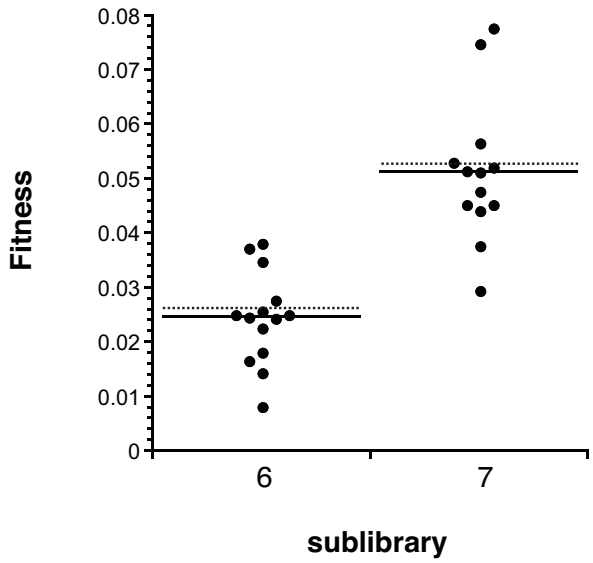


Figure 4.18. Representative western blots from protein dose quantification. The I13E allele (a mutation in the signal sequence) has a higher molecular weight band that corresponds to the size of the protein if the signal sequence has not been removed.

A

Sub-library 6	Sub-library 7
I13E	M68S
E37H	D85R
Y46A	L91R
D50*	I127A
K55P	N136Y
D85I	N136F
I127S	L137N
L137W	L138T
P145Y	G144R
D176R	P145L
T180Q	D176V
T180N	T180I
T181V	R244E
W210S	

B



C

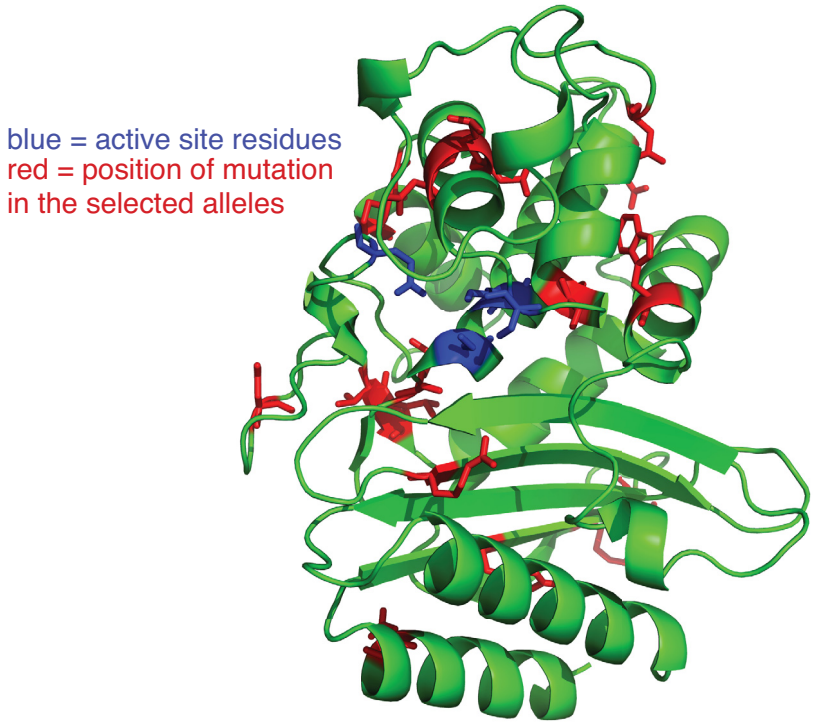


Figure 4.19. Randomly selected members of sub-libraries 6 and 7. These 27 members were selected on plates with 8 μg/ml Amp (sub-library 6) or 16 Amp (sub-library 7). (A) Mutations in the 27 alleles (each allele had one mutation). * indicates the mutation is to the amber stop codon (UAG). (B) Fitness values. The solid line indicates the mean of the randomly selected members. The dotted line indicates the expected mean of the sub-

library based on the Amp concentration used in the genetic selection to obtain the sub-library. (C) The distribution of mutational sites on the structure of TEM-1 (108). Red indicates the mutational sites and blue indicates the four key active site residues (S70, K73, S130 and E166).

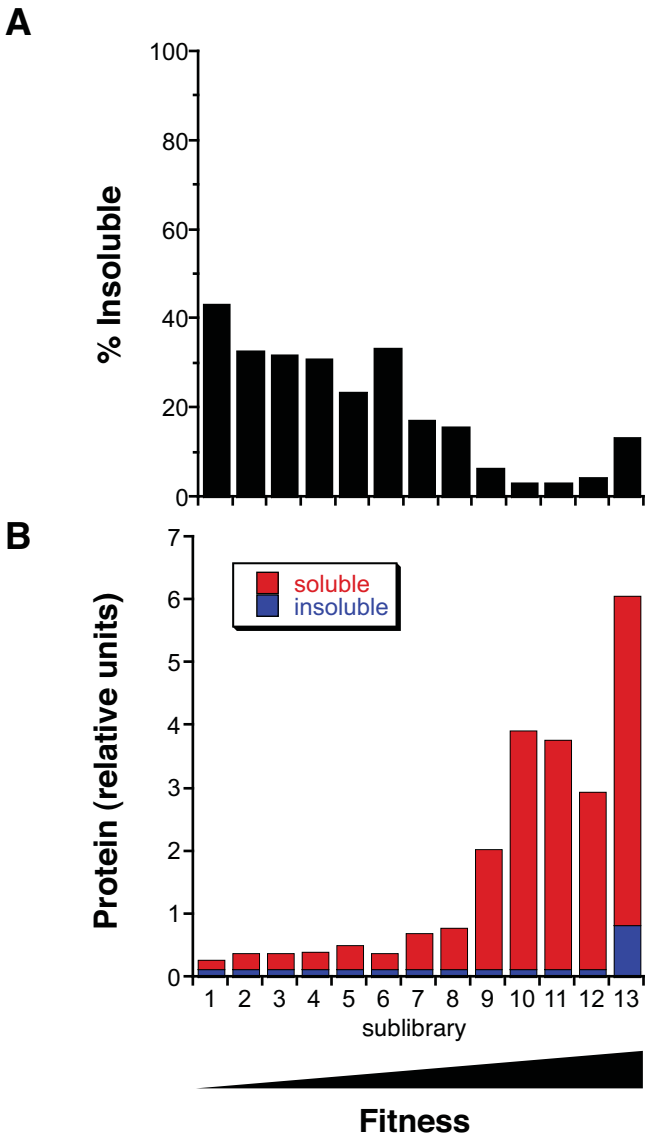


Figure 4.20. A decrease in fitness is not accompanied by an increase in insoluble TEM-1. (A) The percentage of TEM-1 protein in an insoluble state increased as fitness decreased. (B) However, this trend results from a decrease in the soluble TEM-1 and not from an increase in insoluble TEM-1. A representative western blot of the insoluble fraction is shown in Figure 4.18.

We do not interpret the diminished role of mutational effects on protein dose as necessarily reducing the role of thermodynamic stability in fitness. Protein stability, in addition to its effect on protein dose, may exert its effect on fitness through a decrease in a protein's specific activity. Perhaps this manifests by perturbing the conformational ensemble away from more active states or by increasing the number of states (i.e. altering protein dynamics). Protein dose's relative resilience to decreases in thermodynamic stability is striking. We propose that the cellular environment acts as a buffer for deleterious mutational effects on protein dose. The cellular environment may minimize a destabilizing mutation's effect on protein dose, for example, through the effect of chaperones, proteins that facilitate the proper folding of other proteins (143). This theory offers an explanation for TEM-1's stability threshold that buffers the effect of mutations on fitness (76). As such, we suspect that the relative contribution of protein dose to fitness may increase with the number of mutations as the protein's stability margin is exhausted by the cumulative effect of mutations, an effect that is characterized by negative epistasis (76). As such, negative epistasis may arise in part as a consequence of the beneficial properties of the cellular environment in addition to a protein's intrinsic stability margin.

4.4 CONCLUSIONS

The application of synthetic biology to the study of fundamental biological questions, as we have done in this study of fitness landscapes, offers a well-defined, systematic approach for testing and generating hypotheses. Our comprehensive determination of the fitness effects of mutation of *TEM-1* provides the first detailed maps of a fitness

landscape corresponding to a gene and its nearest neighbors at the base pair, codon, and amino acid level. To the extent that *TEM-1* is a representative gene, our study provides several important insights. Evolution must traverse fitness landscapes under the constraints of the genetic code – constraints that minimize the effect of mutation and enrich for adaptive mutations. The small fitness effects of synonymous mutations have complex determinants including regional proclivities for synonymous fitness effects in the gene. At the beginning of the gene, fitness effects of synonymous mutations strongly correlate with mRNA stability. Missense mutational effects on thermodynamic stability shape the DFE, but their deleterious effects on specific protein activity exceed that on protein dose, on average. We hypothesize that *TEM-1*'s high mutational tolerance may in part derive from the cell's buffering capacity to mediate the deleterious effects of lost stability on protein dose, a phenomena that would give rise to negative epistasis. Further inquiry into the fundamental determinants of the landscape's topology is necessary to address this hypothesis and substantiate these findings.

4.5 ACKNOWLEDGMENTS

We thank Yousif Shamoo and Barrett Steinberg for helpful comments on the manuscript. Jason W. Labonte and Jeffrey J. Gray provided expertise concerning the computational prediction of protein thermodynamic stability, and Jason Labonte wrote the PyRosetta script.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

5.1 OPPORTUNITIES AFFORDED BY PFUNKEL MUTAGENESIS

PFunkel offers the opportunity to explore large swathes of protein sequence space in a rapid and relatively inexpensive manner. Comprehensive codon mutagenesis (CCM) libraries of genes, previously unfeasible due to extreme cost and labor requirements, can now be created routinely. Equally importantly, multi-site mutagenesis libraries where multiple codons can be randomized simultaneously are routine using PFunkel. Using these types of PFunkel libraries, we discovered many new alleles of *TEM-1* displaying high activity on several substrates. Many of these variants are unlikely to be found using traditional mutagenesis methodologies. Hence, PFunkel mutagenesis holds promise for evolving improved proteins of all varieties.

From a public health perspective, an extensive knowledge of the possible molecular determinants of bacterial resistance to antibiotics and inhibitors would inform the development and implementation of new antibiotics and inhibitors. We demonstrated the use of PFunkel for discovering many *TEM-1* alleles conferring resistance to tazobactam, a β-lactamase inhibitor used clinically in conjunction with the extended spectrum antibiotic, piperacillin. Some of these alleles confer higher resistance than all known variants, are accessible by point mutation, and have the potential to emerge in the clinic and pose a public health threat. Thus, PFunkel can preemptively identify potentially dangerous resistance mutations that would not otherwise be identified so that new or alternative treatment plans can be developed. Additionally, PFunkel can be used to study

genetic variants of human genes in a comprehensive manner. For example, one could create a comprehensive library of the cancer-associated BRCA1/2 genes in order to study the molecular effects of all such mutations in human cells and their role in cancer progression.

5.1.1 Genetic diversity provided by PFunkel CCM for directed evolution

On-going studies are investigating mutational trajectories that bypass the constraints of the genetic code by using PFunkel CCM to allow all amino acid substitutions. This contrasts with most directed evolution methodologies that utilize error-prone PCR to introduce point mutations as the basis for genetic diversity. CCM serves to smoothen and expand the fitness landscape by providing many more avenues for evolution to traverse, including more beneficial routes. This is especially important when the fitness landscape is rugged, as may be due to epistasis, and it becomes easy to get stuck in local maxima. Using the band-pass system with *TEM-1*, various alternative trajectories are being explored that pass through nodes of low fitness and can therefore reach new areas of sequence space. This directly relates to the question posed by John Maynard Smith, "How often, if ever, has evolution passed through a non-functional sequence (1)?"

5.2 EXPLORING THE DFE AND THE PREVALENCE OF EPISTASIS

Comprehensive DFEs, akin to the one detailed in Chapter 4, can be characterized for other genes and proteins using similar methodologies. Furthermore, the fitness landscape can be expanded beyond all nearest neighbors to variants with multiple mutations. An on-going study is characterizing the DFE of a large set of double-mutants

of *TEM-1*. A goal of this project is to determine the prevalence of epistasis. In this case epistasis refers to genetic interaction - when the effects of mutations are non-independent, as demonstrated in Figure 5.1. While evidence does suggest that epistasis is prevalent in protein evolution (76,144), such a large dataset would be extremely useful for forming improved predictive models for protein adaption and better understanding the wider fitness landscape.

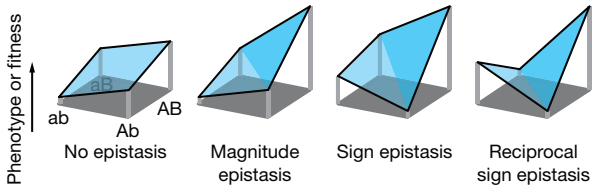


Figure 5.1. Types of epistasis. Considering two mutations from the initial sequence 'ab' to 'AB.' When no epistasis is present, the fitness effects of 'A' and 'B' are independent and additive. In magnitude epistasis the combined effect of 'A' and 'B' is greater than the individual effects but the sign stays the same, while in sign epistasis the sign of the fitness effect changes for one of the mutations. In reciprocal sign epistasis the sign of the fitness effect changes for both mutations (*87*).

5.3 FURTHERING THE EVOLVABILITY THEORY OF THE GENETIC CODE

We have demonstrated using five different data sets that the adaptive set of mutations are enriched in amino acid substitutions accessible by point mutation – evidence that the architecture of the genetic code enriches for adaptive mutations and therefore promotes protein evolvability. We demonstrated this from the DFE of two small influenza inhibitor proteins engineered using computational design and directed evolution. We also demonstrated this property for the TEM-1 protein on the set of adaptive mutations conferring cefotaxime resistance, tazobactam resistance, and ampicillin resistance. Therefore, our evolvability theory holds for a variety of protein types. The influenza inhibitors are binding proteins that are evolutionarily immature and relatively

non-robust to mutation. TEM-1 is a native evolutionarily mature catalytic enzyme that is robust to mutation. The gain-of-function mutations studied are relevant for a varied set of substrates; ampicillin, a first-generation β-lactam antibiotic, cefotaxime, a third-generation cephalosporin antibiotic, and tazobactam, a β-lactamase inhibitor. Though conclusively proving our evolvability theory is difficult, our data provides strong evidence in its favor, and analysis of more DFEs for different proteins will provide greater perspective.

We have reconciled our evolvability theory with the adaptive theory in regards to error-minimization. A code's error minimization must be balanced by its propensity to promote the evolution of proteins. A code maximized for robustness to error would only allow the most conservative mutations, which may not be optimal for protein evolution. Therefore, the evolvability theory provides an explanation for the extent to which, if any, the code is not optimized for error-minimization. On the other hand, it may be the case that the enrichment of adaptive mutations observed is simply a side-effect of error-minimization. An error-minimizing code increases the probability of an adaptive mutation by decreasing the effects of mutations as posited by Fisher's geometric theorem (100). If error-minimization and enrichment of adaptive mutations come together as a package, an interesting but difficult question to address experimentally is the extent to which each contributed to the origin of the genetic code. However, combining experimental DFE data with computational approaches can begin to address these questions.

5.3.1 Combining the experimental DFE with computational calculation of the genetic code's optimality

Computational studies comparing the efficiency of error-minimization for randomly generated genetic codes have shown that the natural genetic code ranks better than the vast majority of codes; one study showed that only one in a million randomly generated codes have higher efficiency (145). The major criticisms leveled against the argument that the genetic code underwent evolutionary optimization for error-minimization stem from computational search algorithms that identify codes with significantly greater error-minimization and that such codes look nothing like the natural code. These criticisms overlook two important considerations. Firstly, these search algorithms are limited by the imperfect and simplified measures of amino acid similarity incorporated into them. Even small changes to these amino acid models can lead to drastically different results. Secondly, these criticisms overlook the role of the evolutionary trajectory the code took, and the codon assignments that were accessible to it due to the mechanisms by which codons could be reassigned. Therefore, as the genetic code adapted, many of the potential genetic codes were not possible (99).

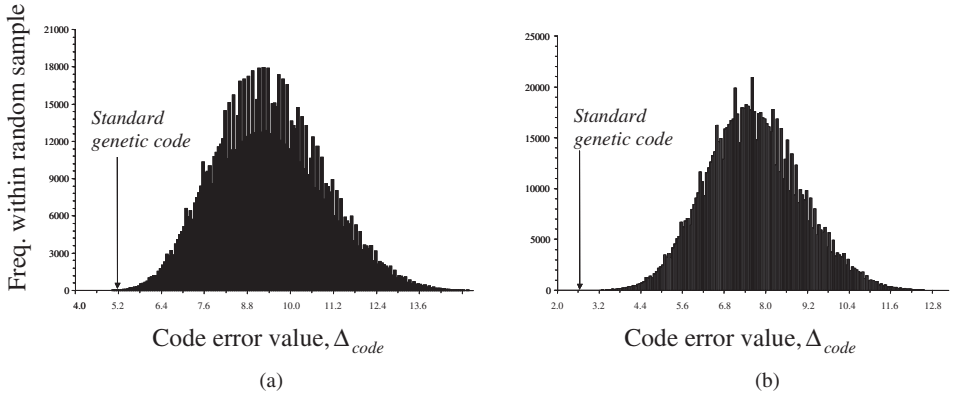


Figure 5.2. Error-minimization comparison for 1 million randomly generated genetic codes. a) Without any transition bias or codon-position weighting, measuring amino acid similarity in terms of hydrophobicity, a proportion of 0.0001 codes within the random sample provide lower error values than the standard genetic code. b) Incorporating experimental patterns of mistranslation reduces the proportion to 0.000001 (*145*).

Comprehensive DFE data offers new avenues to address these questions by combining experimental and computational approaches. Studies can be performed using experimental DFE data together with many if not all possible random configurations of the 64 codon-20 amino acid (plus stop codon) genetic code ($\sim 10^{72}$ total possible arrangements). While computationally demanding, using experimental DFE data overcomes the limitations of previous computational experiments reliant on theoretical models of amino acid similarity (73,99). For each random genetic code, the gene's wildtype and mutated codons would be appropriately reassigned and the error-minimization and adaptive enrichment calculation performed. One could then measure the fraction of genetic codes that provide greater error-minimization (reduce the overall effects of point mutations) than the standard code, and the fraction of codes that provide greater adaptive enrichment than the standard code. Such information would provide insight into the relative contributions of each of these factors in the origin of the genetic code. As more DFE datasets become available, the more useful this calculation would become.

REFERENCES

1. Smith, J. M. Natural selection and the concept of a protein space. *Nature* **225**, 563–564 (1970).

2. Terekhanova, N. V., Bazykin, G. A., Neverov, A., Kondrashov, A. S. & Seplyarskiy, V. B. Prevalence of multinucleotide replacements in evolution of primates and Drosophila. *Mol Biol Evol* **30**, 1315–1325 (2013).

3. Woese, C. R. On the evolution of the genetic code. *Proc. Natl. Acad. Sci. U.S.A.* **54**, 1546–1552 (1965).

4. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).

5. Gamow, G. Possible relation between deoxyribonucleic acid and protein structures. *Nature* **173**, 318 (1954).

6. Koonin, E. V. & Novozhilov, A. S. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**, 99–111 (2009).

7. Sonneborn, T. M.in *Evolving Genes and Proteins* (Bryson, V. & Voge, H. J.) 377–397 (Academic Press, 1965).

8. Woese, C. R., Dugre, D. H., Saxinger, W. C. & Dugre, S. A. The molecular basis for the genetic code. *Proc. Natl. Acad. Sci. U.S.A.* **55**, 966–974 (1966).

9. Crick, F. H. The origin of the genetic code. *Journal of Molecular Biology* **38**, 367–379 (1968).

10. Wong, J. T. A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. U.S.A.* **72**, 1909–1912 (1975).

11. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. in *Proceedings of the sixth international congress of genetics* (1932).

12. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* **10**, 866–876 (2009).

13. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat Rev Genet* **8**, 610–618 (2007).

14. Lind, P. A., Berg, O. G. & Andersson, D. I. Mutational robustness of ribosomal protein genes. *Science* **330**, 825–827 (2010).

15. Peris, J. B., Davis, P., Cuevas, J. M., Nebot, M. R. & Sanjuán, R. Distribution of fitness effects caused by single-nucleotide substitutions in bacteriophage f1. *Genetics* **185**, 603–609 (2010).

16. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function

relationships. *Nat Meth* **7**, 741–746 (2010).

17. Araya, C. L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16858–16863 (2012).

18. Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E1263–72 (2013).

19. Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. A. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J Mol Biol* **425**, 1363–1377 (2013).

20. Whitehead, T. A. *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat Biotech* **30**, 543–548 (2012).

21. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. From the Cover: Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences* **108**, 7896–7901 (2011).

22. Wloch, D. M., Szafraniec, K., Borts, R. H. & Korona, R. Direct estimate of the mutation rate and the distribution of fitness effects in the yeast Saccharomyces cerevisiae. *Genetics* **159**, 441–452 (2001).

23. Deng, Z. *et al.* Deep Sequencing of Systematic Combinatorial Libraries Reveals Beta-Lactamase Sequence Constraints at High Resolution. *J Mol Biol* **424**, 150–167 (2012).

24. McLaughlin, R. N., Jr, Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* 1–7 (2012).doi:10.1038/nature11500

25. Schlinkmann, K. M. *et al.* Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations. *Proceedings of the National Academy of Sciences* **109**, 9810–9815 (2012).

26. Hutchison, C. A. *et al.* Mutagenesis at a specific position in a DNA sequence. *J. Biol. Chem.* **253**, 6551–6560 (1978).

27. Kunkel, T. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 488–492 (1985).

28. Kunkel, T. A., Roberts, J. D. & Zakour, R. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Methods Enzymol* **154**, 367–382 (1987).

29. Braman, J., Papworth, C. & Greener, A. Site-directed mutagenesis using double-

stranded plasmid DNA templates. *Methods Mol Biol* **57**, 31–44 (1996).

30. *QuikChange Site-Directed Mutagenesis Kit Instruction Manual*. (Stratagene).

31. Dominy, C. N. & Andrews, D. W. Site-directed mutagenesis by inverse PCR. *Methods Mol Biol* **235**, 209–223 (2003).

32. Bi, W. & Stambrook, P. J. Site-directed mutagenesis by combined chain reaction. *Anal Biochem* **256**, 137–140 (1998).

33. McCullum, E. O., Williams, B. A. R., Zhang, J. & Chaput, J. C. Random mutagenesis by error-prone PCR. *Methods Mol Biol* **634**, 103–109 (2010).

34. Baldwin, A. J., Busse, K., Simm, A. M. & Jones, D. D. Expanded molecular diversity generation during directed evolution by trinucleotide exchange (TriNEx). *Nucleic Acids Research* **36**, e77–e77 (2008).

35. Murakami, H., Hohsaka, T. & Sisido, M. Random insertion and deletion of arbitrary number of bases for codon-based random mutation of DNAs. *Nat Biotechnol* **20**, 76–81 (2002).

36. Liu, J. & Cropp, T. A. A method for multi-codon scanning mutagenesis of proteins based on asymmetric transposons. *Protein Engineering Design and Selection* **25**, 67–72 (2012).

37. Hames, C., Halbedel, S., Schilling, O. & St u lke, J. O. R. Multiple-mutation reaction: a method for simultaneous introduction of multiple mutations into the glpK gene of Mycoplasma pneumoniae. *Appl Environ Microbiol* **71**, 4097–4100 (2005).

38. Scholle, M. D., Kehoe, J. W. & Kay, B. K. Efficient construction of a large collection of phage-displayed combinatorial peptide libraries. *Comb. Chem. High Throughput Screen.* **8**, 545–551 (2005).

39. Weiss, G. A., Watanabe, C. K., Zhong, A., Goddard, A. & Sidhu, S. S. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 8950–8954 (2000).

40. Sambrook, J., MacCallum, P. & Russell, D. *Molecular Cloning: A Laboratory Manual*. (Cold Spring Harbor Laboratory Press, 2001).

41. Chung, C. T., Niemela, S. L. & Miller, R. H. One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2172–2175 (1989).

42. Sohka, T. *et al.* An externally tunable bacterial band-pass filter. *Proceedings of the National Academy of Sciences* **106**, 10135–10140 (2009).

43. Trower, M. K. Site-directed mutagenesis using a uracil-containing phagemid template. *Methods Mol Biol* **31**, 67–77 (1994).

44. *PfuTurbo Cx hotstart DNA polymerase Intruction Manual*. (Agilent Technologies, 2009).

45. Goecks, J., Nekrutenko, A., Taylor, J. & Team, G. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**, R86 (2010).

46. Blankenberg, D. *et al.* Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* **Chapter 19**, Unit 19.10.1–21 (2010).

47. Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451–1455 (2005).

48. Wassman, C. D. Predicting oligonucleotide-directed mutagenesis failures in protein engineering. *Nucleic Acids Research* **32**, 6407–6413 (2004).

49. Kunkel, T. A., Bebenek, K. & McClary, J. Efficient site-directed mutagenesis using uracil-containing DNA. *Methods Enzymol* **204**, 125–139 (1991).

50. Taylor, A. F. & Weiss, B. Role of exonuclease III in the base excision repair of uracil-containing DNA. *Journal of Bacteriology* **151**, 351–357 (1982).

51. Nørholm, M. H. A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. *BMC Biotechnol* **10**, 21 (2010).

52. Hogrefe, H. H., Cline, J., Lovejoy, A. E. & Nielson, K. B. DNA polymerases from hyperthermophiles. *Methods Enzymol* **334**, 91–116 (2001).

53. Rogers, S. G. & Weiss, B. Exonuclease III of Escherichia coli K-12, an AP endonuclease. *Methods Enzymol* **65**, 201–211 (1980).

54. *QuikChange Multi Site-Directed Mutagenesis Kit Instruction Manual*. (Stratagene).

55. Denbigh, K. G. Velocity and yield in continuous reaction systems. *Trans. Faraday Soc.* **40**, 352 (1944).

56. Holland, M. M., McQuillan, M. R. & O'Hanlon, K. A. Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy. *Croat Med J* **52**, 299–313 (2011).

57. Meyerhans, A., Vartanian, J. P. & Wain-Hobson, S. DNA recombination during PCR. *Nucleic Acids Research* **18**, 1687–1691 (1990).

58. Lindahl, T. & Nyberg, B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**, 3405–3410 (1974).

59. Lindahl, T. & Nyberg, B. Rate of depurination of native deoxyribonucleic acid. *Biochemistry* **11**, 3610–3618 (1972).

60. André, P., Kim, A., Khrapko, K. & Thilly, W. G. Fidelity and mutational

spectrum of Pfu DNA polymerase on a human mitochondrial DNA sequence. *Genome Res* **7**, 843–852 (1997).

61. Ambler, R. P. & Coulson, F. W. A Standard Numbering Scheme for the Class A Beta-Lactamases. *Biochemical Journal Letters* **276**, 269–272 (1991).

62. Drawz, S. M. & Bonomo, R. A. Three decades of beta-lactamase inhibitors. *Clin Microbiol Rev* **23**, 160–201 (2010).

63. Robin, F. *et al.* In vitro efficiency of the piperacillin/tazobactam combination against inhibitor-resistant TEM- and complex mutant TEM-producing clinical strains of Escherichia coli. *Journal of Antimicrobial Chemotherapy* **66**, 1052–1056 (2011).

64. Vakulenko, S. B., Geryk, B., Kotra, L. P., Mobashery, S. & Lerner, S. A. Selection and characterization of beta-lactam-beta-lactamase inactivator-resistant mutants following PCR mutagenesis of the TEM-1 beta-lactamase gene. *Antimicrobial Agents and Chemotherapy* **42**, 1542–1548 (1998).

65. Cunningham, B. C. & Wells, J. A. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* **244**, 1081–1085 (1989).

66. Araya, C. L. & Fowler, D. M. Deep mutational scanning: assessing protein function on a massive scale. *Trends in Biotechnology* **29**, 435–442 (2011).

67. Leung, D. W., Chen, E. & Goeddel, D. V. A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique* **1**, 11–15 (1989).

68. Burch, C. L. & Chao, L. Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature* **406**, 625–628 (2000).

69. Cambray, G. & Mazel, D. Synonymous genes explore different evolutionary landscapes. *PLoS Genet* **4**, e1000256 (2008).

70. Hall, A. R., Griffiths, V. F., MacLean, R. C. & Colegrave, N. Mutational neighbourhood and mutation supply rate constrain adaptation in Pseudomonas aeruginosa. *Proceedings of the Royal Society B: Biological Sciences* **277**, 643–650 (2010).

71. Fitch, W. M. An improved method of testing for evolutionary homology. *Journal of Molecular Biology* **16**, 9–16 (1966).

72. Stoltzfus, A. & Yampolsky, L. Y. Climbing mount probable: mutation as a cause of nonrandomness in evolution. *J. Hered.* **100**, 637–647 (2009).

73. Zhu, W. & Freeland, S. The standard genetic code enhances adaptive evolution of proteins. *Journal of Theoretical Biology* **239**, 63–70 (2006).

74. Firnberg, E. & Ostermeier, M. PFunkel: efficient, expansive, user-defined mutagenesis. *PLoS ONE* **7**, e52031 (2012).

75. Weinreich, D. M. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science* **312**, 111–114 (2006).

76. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).

77. DePristo, M. A., Hartl, D. L. & Weinreich, D. M. Mutational reversions during adaptive protein evolution. *Molecular Biology and Evolution* **24**, 1608–1610 (2007).

78. Salverda, M. L. M. *et al.* Initial Mutations Direct Alternative Pathways of Protein Evolution. *PLoS Genet* **7**, e1001321 (2011).

79. Poyart, C., Mugnier, P., Quesne, G., Berche, P. & Trieu-Cuot, P. A novel extended-spectrum TEM-type beta-lactamase (TEM-52) associated with decreased susceptibility to moxalactam in Klebsiella pneumoniae. *Antimicrobial Agents and Chemotherapy* **42**, 108–113 (1998).

80. Barlow, M. & Hall, B. G. Predicting evolutionary potential: in vitro evolution accurately reproduces natural evolution of the tem beta-lactamase. *Genetics* **160**, 823–832 (2002).

81. Kopsidas, G. *et al.* RNA mutagenesis yields highly diverse mRNA libraries for in vitro protein evolution. *BMC Biotechnol* **7**, 18 (2007).

82. Orencia, M. C., Yoon, J. S., Ness, J. E., Stemmer, W. P. & Stevens, R. C. Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. *Nat. Struct. Biol.* **8**, 238–242 (2001).

83. Stemmer, W. P. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**, 389–391 (1994).

84. Zaccolo, M. & Gherardi, E. The effect of high-frequency random mutagenesis on in vitro protein evolution: a study on TEM-1 β-lactamase. *Journal of Molecular Biology* **285**, 775–783 (1999).

85. Hayes, R. J. *et al.* Combining computational and experimental screening for rapid optimization of protein properties. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 15926–15931 (2002).

86. Yi, H. *et al.* Twelve Positions in a β-Lactamase That Can Expand Its Substrate Spectrum with a Single Amino Acid Substitution. *PLoS ONE* **7**, e37585 (2012).

87. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**, 383–386 (2007).

88. Cantu, C. & Palzkill, T. The role of residue 238 of TEM-1 beta-lactamase in the hydrolysis of extended-spectrum antibiotics. *J. Biol. Chem.* **273**, 26603–26609 (1998).

89. Shafikhani, S., Siegel, R. A., Ferrari, E. & Schellenberger, V. Generation of large libraries of random mutants in Bacillus subtilis by PCR-based plasmid multimerization. *Biotech.* **23**, 304–310 (1997).

90. Lee, H., Popodi, E., Tang, H. & Foster, P. L. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E2774–83 (2012).

91. Fleishman, S. J. *et al.* Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. *Science* **332**, 816–821 (2011).

92. Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18**, 1368–1373 (2005).

93. Schenk, M. F., Szendro, I. G., Krug, J. & de Visser, J. A. G. M. Quantifying the Adaptive Potential of an Antibiotic Resistance Enzyme. *PLoS Genet* **8**, e1002783 (2012).

94. Salverda, M. L. M., de Visser, J. A. G. M. & Barlow, M. Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS Microbiology Reviews* **34**, 1015–1036 (2010).

95. Gould, S. J. The evolutionary biology of constraint. *Daedalus* **109**, 39 (1980).

96. Colegrave, N. & Collins, S. Experimental evolution: experimental evolution and evolvability. *Heredity (Edinb)* **100**, 464–470 (2008).

97. Sniegowski, P. D. & Murphy, H. A. Evolvability. *Curr. Biol.* **16**, R831–4 (2006).

98. Lynch, M. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl. Acad. Sci. U.S.A.* **104 Suppl 1**, 8597–8604 (2007).

99. Freeland, S. J. The Darwinian genetic code: an adaptation for adapting? *Genetic Programming and Evolvable Machines* **3**, 113–127 (2002).

100. Fisher, R. A. *The Genetical Theory of Natural Selection*. (The Clarendon Press, Oxford, 1930).

101. Orr, H. A. The genetic theory of adaptation: a brief history. *Nat Rev Genet* **6**, 119–127 (2005).

102. Dietz, H., Pfeifle, D. & Wiedemann, B. The signal molecule for beta-lactamase induction in Enterobacter cloacae is the anhydromuramyl-pentapeptide. *Antimicrobial Agents and Chemotherapy* **41**, 2113–2120 (1997).

103. Valtonen, S. J., Kurittu, J. S. & Karp, M. T. A Luminescent Escherichia coli Biosensor for the High Throughput Detection of -Lactams. *Journal of*

Biomolecular Screening **7**, 127–134 (2002).

104. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nature Publishing Group* **12**, 32–42 (2010).

105. Hofacker, I. L. RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc Bioinformatics* **Chapter 12**, Unit12.2 (2009).

106. Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z. & Blüthgen, N. Efficient translation initiation dictates codon usage at gene start. *Molecular Systems Biology* **9**, 675 (2013).

107. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).

108. Fonzé, E. *et al.* TEM1 beta-lactamase structure solved by molecular replacement and refined structure of the S235A mutant. *Acta Crystallogr. D Biol. Crystallogr.* **51**, 682–694 (1995).

109. Das, R. & Baker, D. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77**, 363–382 (2008).

110. Shapovalov, M. V. & Dunbrack, R. L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**, 844–858 (2011).

111. Dehouck, Y., Kwasigroch, J. M., Gilis, D. & Rooman, M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics* **12**, 151 (2011).

112. Soskine, M. & Tawfik, D. S. Mutational effects and the evolution of new protein functions. *Nat Rev Genet* **11**, 572–582 (2010).

113. Jacquier, H. *et al.* Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 13067–13072 (2013).

114. Camps, M., Herman, A., Loh, E. & Loeb, L. A. Genetic Constraints on Protein Evolution. *Critical Reviews in Biochemistry and Molecular Biology* **42**, 313–326 (2007).

115. Sanju a n, R., Moya, A. E. S. & Elena, S. F. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 8396–8401 (2004).

116. Firnberg, E. & Ostermeier, M. The genetic code constrains yet facilitates Darwinian evolution. *Nucleic Acids Res* (2013).doi:10.1093/nar/gkt536

117. Hilterbrand, A., Saelens, J. & Putonti, C. CBDB: the codon bias database. *BMC Bioinformatics* **13**, 62 (2012).

118. Bossi, L. Context effects: translation of UAG codon by suppressor tRNA is affected by the sequence following UAG in the message. *Journal of Molecular Biology* **164**, 73–87 (1983).

119. Miller, J. H. & Albertini, A. M. Effects of surrounding sequence on the suppression of nonsense codons. *Journal of Molecular Biology* **164**, 59–71 (1983).

120. Singaravelan, B., Roshini, B. R. & Munavar, M. H. Evidence that the supE44 mutation of Escherichia coli is an amber suppressor allele of glnX and that it also suppresses ochre and opal nonsense mutations. *J Bacteriol* **192**, 6039–6044 (2010).

121. Sacerdot, C. *et al.* The role of the AUU initiation codon in the negative feedback regulation of the gene for translation initiation factor IF3 in Escherichia coli. *Mol Microbiol* **21**, 331–346 (1996).

122. Sussman, J. K., Simons, E. L. & Simons, R. W. Escherichia coli translation initiation factor 3 discriminates the initiation codon in vivo. *Mol Microbiol* **21**, 347–360 (1996).

123. Haggerty, T. J. & Lovett, S. T. IF3-mediated suppression of a GUA initiation codon mutation in the recJ gene of Escherichia coli. *Journal of Bacteriology* **179**, 6705–6713 (1997).

124. Gierasch, L. M. Signal sequences. *Biochemistry* **28**, 923–930 (1989).

125. Sougakoff, W. *et al.* Characterization of the plasmid genes blaT-4 and blaT-5 which encode the broad-spectrum beta-lactamases TEM-4 and TEM-5 in enterobacteriaceae. *Gene* **78**, 339–348 (1989).

126. Shenkin, P. S., Erman, B. & Mastrandrea, L. D. Information-theoretical entropy as a measure of sequence variability. *Proteins* **11**, 297–313 (1991).

127. Walkiewicz, K. *et al.* Small changes in enzyme function can lead to surprisingly large fitness effects during adaptive evolution of antibiotic resistance. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 21408–21413 (2012).

128. Couñago, R., Wilson, C. J., Peña, M. I., Wittung-Stafshede, P. & Shamoo, Y. An adaptive mutation in adenylate kinase that increases organismal fitness is linked to stability-activity trade-offs. *Protein Eng Des Sel* **21**, 19–27 (2008).

129. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. & Tawfik, D. S. The stability effects of protein mutations appear to be universally distributed. *Journal of Molecular Biology* **369**, 1318–1332 (2007).

130. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* **6**, 678–687 (2005).

131. Wylie, C. S. & Shakhnovich, E. I. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proceedings of the National Academy of Sciences* **108**, 9916–9921 (2011).

132. Pakula, A. A., Young, V. B. & Sauer, R. T. Bacteriophage lambda cro mutations: effects on activity and intracellular degradation. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8829–8833 (1986).

133. Schultz, S. C. & Richards, J. H. Site-saturation studies of beta-lactamase: production and characterization of mutant beta-lactamases with all possible amino acid substitutions at residue 71. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 1588–1592 (1986).

134. Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 452–456 (1995).

135. Wang, X., Minasov, G. & Shoichet, B. K. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *Journal of Molecular Biology* **320**, 85–95 (2002).

136. Potapov, V., Cohen, M. & Schreiber, G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Engineering Design and Selection* **22**, 553–560 (2009).

137. Brown, N. G., Pennington, J. M., Huang, W., Ayvaz, T. & Palzkill, T. Multiple global suppressors of protein stability defects facilitate the evolution of extended-spectrum TEM β-lactamases. *J Mol Biol* **404**, 832–846 (2010).

138. Kather, I., Jakob, R. P., Dobbek, H. & Schmid, F. X. Increased folding stability of TEM-1 beta-lactamase by in vitro selection. *J Mol Biol* **383**, 238–251 (2008).

139. Raquet, X. *et al.* Stability of TEM beta-lactamase mutants hydrolyzing third generation cephalosporins. *Proteins* **23**, 63–72 (1995).

140. Guillaume, G. *et al.* Site-directed mutagenesis of glutamate 166 in two beta-lactamases. Kinetic and molecular modeling studies. *J. Biol. Chem.* **272**, 5438–5444 (1997).

141. Wang, X., Minasov, G. & Shoichet, B. K. Noncovalent interaction energies in covalent complexes: TEM-1 beta-lactamase and beta-lactams. *Proteins* **47**, 86–96 (2002).

142. Bershtein, S., Goldin, K. & Tawfik, D. S. Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins. *Journal of Molecular Biology* **379**, 1029–1044 (2008).

143. Rutherford, S. L. Between genotype and phenotype: protein chaperones and evolvability. *Nat Rev Genet* **4**, 263–274 (2003).

144. Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535–538 (2012).

145. Freeland, S. J. & Hurst, L. D. The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248 (1998).

APPENDIX 1 – STRAINS

For a full list of strains and constructs including all TEM-1 mutants see Ostermeier lab wiki page.

Name	E. Coli Strain	Plasmid 1	Resistance 1	Plasmid 2	Resistance 2	Plasmid 3	Resistance 3	Date	Comments
JM109	JM109	-	-	-	-	-	-	-	-
SNO301	SNO301	-	Strep 50 ug/mL (but not Spec)	-	-	-	-	-	-
CJ236	CJ236	-	Cm 25 ug/mL (on F episome)	-	-	-	-	-	-
XL-1 Blue	XL-1 Blue	-	Tet 20 ug/mL	-	-	-	-	-	-
DH5α	DH5α	-	Nalidixic Acid 30 ug/mL	-	-	-	-	-	-
ES1301	ES1301	-	-	-	-	-	-	-	-
DH10B	DH10B	-	Strep 50 ug/mL	-	-	-	-	-	-
EF01	RH11	pTS1	Spec 50 ug/mL	pTHY56	Cm 50 ug/mL	-	-	-	-
EF02	DH5α	pTHY56	Cm 50 ug/mL	-	-	-	-	-	-
EF03	RH12	pEF01	Cm 50 ug/mL	pTS1	Spec 50 ug/mL	-	-	-	-
EF04	RH12	pEF02	Cm 50 ug/mL	pTS1	Spec 50 ug/mL	-	-	-	-
EF05	CJ236	pRSET A	Amp 100ug/mL	Cm 25 ug/mL (on F episome)	-	-	-	-	-
EF06	RH12	pTS1	Spec 50 ug/mL	pBAD18-ATB5	Kan 50 ug/mL	-	-	-	-
EF07	RH12	pTS1	Spec 50 ug/mL	pBAD18-kan-ATB5	Kan 50 ug/mL	pDIMC8-cpBLA-LE-AT1	Cm 50 ug/mL	-	cpBLA without peptide, only ATB5 binding site
EF08	RH12	pTS1	Spec 50 ug/mL	pDIMC8-cpBLA-lib3-col1-MICamp32	Cm 50 ug/mL	-	-	-	-
EF09	RH12	pTS1	Spec 50 ug/mL	pDIMC8-cpBLA-lib3-col2-MICamp32	Cm 50 ug/mL	-	-	-	-
EF10	RH12	pTS1	Spec 50 ug/mL	pDIMC8-cpBLA-lib3-col3-MICamp32	Cm 50 ug/mL	-	-	-	-
EF11	RH12	pTS1	Spec 50 ug/mL	pDIMC8-cpBLA-lib3-col1-MICamp2	Cm 50 ug/mL	-	-	-	-
EF12	RH12	pTS1	Spec 50 ug/mL	pDIMC8-cpBLA-lib3-col1-MICamp4	Cm 50 ug/mL	-	-	-	-
EF13	RH12	pTS1	Spec 50 ug/mL	pDIMC8-cpBLA-lib3-col1-MICamp8	Cm 50 ug/mL	-	-	-	-
EF14	RH12	pTS1	Spec 50 ug/mL	pDIMC8-cpBLA-lib3-col1-MICamp16	Cm 50 ug/mL	-	-	-	-
EF15	RH12	pTS1	Spec 50 ug/mL	pDIMC8-cpBLA-lib3-col1-MICamp32	Cm 50 ug/mL	-	-	-	-
EF16	DH5α	pDIMC8-cpBLA-lib4-C1A0 (from cpBLA-lib4-C1A0)	Cm 50 ug/mL	-	-	-	-	1/10/10	-

EF17	EF06	pTS1(from cpBLA-lib4-C3A0)	Spec 50 ug/mL	pBAD18-kan-ATB5 (from cpBLA-lib4-C3A0)	Kan 50 ug/mL	pDIMC8-cpBLA-lib4-C3A0 (from cpBLA-lib4-C3A0)	Cm 50 ug/mL	1/10/10	constitutive TetC expression
EF18	DH5α	pBAD18-kan-ATB5 (from cpBLA-lib4-C1A0)	Kan 50 ug/mL	-	-		-	1/10/10	-
EF19	DH5α	pTS1(from cpBLA-lib4-C1A0)	Spec 50 ug/mL	-	-		-	-	-
EF20	EF06	pTS1(from cpBLA-lib4-C4A0)	Spec 50 ug/mL	pBAD18-kan-ATB5 (from cpBLA-lib4-C4A0)	Kan 50 ug/mL	pDIMC8-cpBLA-lib4-C4A0 (from cpBLA-lib4-C4A0)	Cm 50 ug/mL	1/14/10	constitutive TetC expression
EF21	DH5α	pBAD18-kan-ATB5 (original)	Kan 50 ug/mL	-	-	-	-	-	-
EF22	RH12	pTS1	Spec 50 ug/mL	pBAD18-kan-ATB5	Kan 50 ug/mL	pDIMC8-cpBLA-lib4-C85A1	Cm 50 ug/mL	2/10/10	library member which should grow on Amp 16, 32, 64 ug/mL band-pass plates
EF23	CJ236	pSkunk1-TEM1bla	Spec 50 ug/mL	Cm 25 ug/mL (on F episome)	-	-	-	-	-
EF24	ER2420	pACYC184	Cm 50ug/mL, Tet 30 ug/mL	-	-	-	-	-	strain from NEB #E4125S
EF25	RH12	pTS1	Spec 50 ug/mL	pTHY56	Cm 50 ug/mL	pBAD18-kan-ATB5	Kan 50 ug/mL		-
EF26	DH5α	pTS38	Spec 50 ug/mL	-	-	-	-	-	-
EF27	SNO301	pTS38	Spec 50 ug/mL	-	-	-	-	-	-
EF28	SNO301	pTS38	Spec 50 ug/mL	pDIMC8-BLA	Cm 50 ug/mL	-	-	-	-
EF29	SNO301	pBAD-kana-mod1-ATB5 (old, not good, bad RBS spacing)	Kan 50 ug/mL	-	-	-	-	-	-
EF30	DH5α	pBAD-kana-mod1-ATB5 (old, not good, bad RBS spacing)	Kan 50 ug/mL	-	-	-	-	-	-
EF31	DH5α	pBAD-kana-mod1-ATB5 (new, expression confirmed)	Kan 50 ug/mL	-	-	-	-	5/10/10	-
EF32	DH5α	pDIMC8-cpBLA-LE-AT1	Cm 50 ug/mL	-	-	-	-	-	-
EF33	SNO301	pTS39	Kan 50 ug/mL	-	-	-	-	-	-
EF34	SNO301	pTS39	Kan 50 ug/mL	pTHY56	Cm 50 ug/mL	-	-	-	-
EF35	SNO301	pTS39	Kan 50 ug/mL	pDIMC8-cpBLA-ATI	Cm 50 ug/mL	-	-	-	-
EF36	SNO301	pTS40	Cm 50 ug/mL	-	-	-	-	-	-
EF37	DH5α	pDIMC8-TEM1bla	Cm 50 ug/mL	-	-	-	-	-	-
EF38	SN0301	PTS1	Spec 50 ug/mL	pDIMC8-cpBLA-ATI	Cm 50 ug/mL	-	-	-	-
EF39	DH5α	pBAD-kana-mod1-ATB5 (new, expression confirmed)	Kan 50 ug/mL	pDIMC8-cpBLA-ATI	Cm 50 ug/mL	-	-	-	-
EF40	DH5α	pBR322	Tet 20 ug/mL, Amp 100ug/mL	-	-	-	-	-	-
EF41	DH5α	pM13helperplasmid	Tet 20 ug/mL	-	-	-	-	-	-
EF42	DH5α	pEF03	Amp 100ug/mL	-	-	-	-	-	-
EF43	SN0301	pTS40	Cm 50 ug/mL	pEF03	Amp 100ug/mL	-	-	-	-

EF44	SNO301	pTS40	Cm 50 ug/mL	pSkunk2-natBLA	Spec 50 ug/mL	-	-	-	-
EF45	DH5α	pSkunk1-TEM1bla	Spec 50 ug/mL	-	-	-	-	-	-
EF46	SNO301	pTS40	Cm 50 ug/mL	pSkunk1-TEM1bla	Spec 50 ug/mL	-	-	-	-
EF47	DH5α	pSkunk2-TEM1bla	Spec 50 ug/mL	-	-	-	-	-	-
EF48	SNO301	pTS40	Cm 50 ug/mL	pSkunk2-TEM1bla	Spec 50 ug/mL	-	-	-	-
EF49	DH5α	pTS42	Cm 50 ug/mL	-	-	-	-	-	-
EF50	DH5α	pSkunk3-TEM1bla	Spec 50 ug/mL	-	-	-	-	-	-
EF51	SNO301	pTS42	Cm 50 ug/mL	-	-	-	-	-	-
EF52	SN0301	pTS42	Cm 50 ug/mL	pSkunk2-TEM1bla	Spec 50 ug/mL	-	-	-	-
EF53	SNO301	pTS42	Cm 50 ug/mL	pSkunk3-TEM1bla	Spec 50 ug/mL	-	-	-	-
EF54	CJ236	pSkunk3-TEM1bla	Spec 50 ug/mL	-	-	-	-	-	Cm 15 ug/mL for F episome, propogate at 30C in LB +Thd 125 ug/mL
EF55	DH5α	pSkunk3-QuadBLA	Spec 50 ug/mL	-	-	-	-	-	-
EF56	SN0301	pTS42	Cm 50 ug/mL	pSkunk3-QuadBLA	Spec 50 ug/mL	-	-	-	-
EF57	CJ236	pEF03	Amp 100ug/mL	-	-	-	-	-	Cm 15 ug/mL for F episome, propogate at 30C in LB +Thd 125 ug/mL
EF58	CJ236	pSkunk2-natBLA	Spec 50 ug/mL	-	-	-	-	-	Cm 15 ug/mL for F episome, propogate at 30C in LB +Thd 125 ug/mL
EF59	NEB 5-alpha F'Iq	pSkunk3-TEM1bla	Spec 50 ug/mL	-	-	-	-	-	Tet 15 ug/mL for F episome
EF60	SN0301	pSkunk3-TEM1bla-stop	Spec 50 ug/mL	pTS40	Cm 50 ug/mL	-	-	-	-
EF61	SNO301	pTS42	Cm 50 ug/mL	pSkunk2-MCS	Spec 50 ug/mL	-	-	-	-

APPENDIX II – PLASMIDS

For a full list of plasmids, including all *TEM-1* mutants, see Ostermeier lab wiki page.

Name	Vector	Genes	Resistance	Date	Comments
pTHY56	pDIMC8-Mlu1-SD	cpBLA	Cm 50 ug/mL		from Tae Hyeon Yoo at Georgiou lab
pEF01	pDIMC8	cpBLA-lib3-amp2-peptide1	Cm 50 ug/mL		
pEF02	pDIMC8	cpBLA-lib3-amp2-peptide2	Cm 50 ug/mL		
pBAD18-kan-ATB5	pBAD18-kan	ATB5 scFv against human angiotensin (AT I)	Kan 50 ug/mL	10/5/09	from Tae Hyeon Georgiou Lab (includes sequences of TEM1 BLA)
pSkunk1-BLA	pSkunk1	TEM1-BLA	Spec 50 ug/mL		pDIMC8 with Cm resistance swapped with Strep
pTS38	pTS1	AmpR, TetC (D17N), GFP	Spec 50 ug/mL	3/1/10	same as pTS1 only with D17N mutation in TetC
pBAD-kana-mod1-ATB5 (old, bad RBS spacing)	pBAD-kana-mod1	ATB5 scFv against human angiotensin (AT I)	Kan 50 ug/mL	3/30/10	ATB5 gene inserted into Taka's pBAD-kana-mod1 vector (lacks BLA sequences)
pBAD-kana-mod1-ATB5 (new, tested expression ok)	pBAD-kana-mod1	ATB5 scFv against human angiotensin (AT I)	Kan 50 ug/mL	3/30/10	ATB5 gene cassette from pBAD18-kan-ATB5 inserted into Taka's pBAD-kana-mod1 vector (lacks BLA sequences)
pDIMC8-cpBLA-LE-AT1	pDIMC8	cpBLA with ATI antibody binding site	Cm 50 ug/mL		
pTS39	pTS1	pTS1 with AraC-pBAD promoter-ATB5-Kan cassette from pBAD-kana-mod1-ATB5-new	Kan 50 ug/mL	5/17/10	
pTS40	pTS1	pTS1 w Cam resistance	Cm 50 ug/mL	8/5/10	
M13HP (helper plasmid)	M13KO7	insert from pBR322 from TetR to ColE1 origin	Tet 20 ug/mL	7/18/11	M13 origin & KanR removed from M13KO7 and TetR and ColE1 origin from pBR322 inserted
pEF03	pBR322	removed TetR and inserted f1 ori	Amp 100 ug/mL	9/8/11	For usage in band-pass, has BLA with natural promoter.
pSkunk2-natBLA	pSkunk2-MCS	TEM1-BLA with pBR322 natural promoter	Spec 50 ug/mL	9/8/11	
pSkunk2-TEM1bla	pSkunk2-MCS	TEM1-BLA with Tac promoter	Spec 50 ug/mL	9/20/11	
pSkunk3-TEM1bla	pDIMC8-BLA	TEM1-bla, CmR CDS in pDIMC8 replaced with SmR CDS	Spec 50 ug/mL	9/28/11	
pTS42	pTS1	SmR CDS in pTS1 replaced with CmR CDS	Cm 50 ug/mL	9/28/11	

APPENDIX III – PRIMERS

For a full list of primers, including all *TEM-1* PFunkel mutagenesis primers and 454 deep sequencing adapter primers, please see Ostermeier lab wiki page.

Label	Function	Template	Sequence
300	TAC for	pDIMC8 before cloned gene	tgacaattaatcatcggctc
301	317-rev (Not as good as P303)	pDIMC8 after cloned gene	cgaccccgaacgccagcaag
302	-200 TAC for	pDIMC8 200 BP before TAC primer	gaagacctgaccgccgcagag
303	317-rev2 (better than P301)	Taka creation - slightly before 317-rev on 3' strand	aaagcaaattcgaccccgaa
304	T7 for (from Clay)	any plasmid with T7 promoter	taatacgactcactataggg
305	BGH (for?)	any plasmid with BGH region	not sure of exact sequence
306	MBP rev SUC OVR	any MBP switch	tctttcgccaactcttcc
307	pSkunk1-seq-rev	pSkunk1 after gene rev	gtttattgactaccggaagcagtgtg
308	pSkunk1-seq-for	pSkunk1 after gene for	tttaacgaccctgccctgaacc
309	QuikChange Tet D17N for	pTS1	ggcaccgtcaccctgaacgctgtaggcataggc
310	QuikChange Tet D17N rev	pTS1	gcctatgcctacagcgttcagggtgacggtgcc
311	pBAD-kana-mod1-for	pBAD start of AraC	P-ttatgacaacttgacggctacatc
312	pBAD-kana-mod1-rev	pBAD end of Kan gene	P-atgagccatattcaacgggaaacg
313	pTS1 after Strep for	pTS1 after Strep for	actcttccttttttcaatattattgaagc
314	pTS1 before Strep rev	pTS1 before Strep rev	ctaacaattcgttcaagccgaggg
315	pBAD promoter for	pBAD18-kan-ATB5	attagcggatcctacctgacgctt
316	pBAD terminator rev	pBAD18-kan-ATB5	cttcgcaacgttcaaatccg
317	Cm for adding KpnI site	pDIMC8 Cm gene	ggtaccttacgccccgccctgccactcat
318	Cm rev adding PstI site	pDIMC8 Cm gene	ctgcaggttccaactttcaccataat
319	pSkunk1-rev (works with pTS40)	pSkunk1 after gene rev (170 bp after gene)	ggtagtcggcaaataatgtc
320	pSkunk3-rev	pSkunk3 after gene (a little after P301)	P-gcagaaattcgaaagcaaattcgac
321	pSkunk p15a origin rev	pSkunk3 after origin rev	gaggttttcaccgtcatc
322	pSkunk3 ori for	pSkunk3 before origin for	agcgctagcggagtgta
323	glnV for	chromosomal tRNA - glnV for	atttaggttctggcgccgcaa
324	glnV rev	chromosomal tRNA - glnV rev	tatggatctgttttatgccggattc
325	leuX for	chromosomal tRNA - leuX for	caatcacctcatgcccggcaa
326	leuX rev	chromosomal tRNA - leuX rev	tcgtcacgacgtactctggcat
327	serU for	chromosomal tRNA - serU for	catcatcgaccacaaatggcg
328	serU rev	chromosomal tRNA - serU rev	acttgttagtatgatacaaaggc

329	trpT for	chromosomal tRNA - trpT for	gcagcgatgcgttgagctaaccg
330	trpT rev	chromosomal tRNA - trpT rev	aattatctgcgacttgagtg
331	tyrT for	chromosomal tRNA - tyrT for	tatcaccaggcctatgaacg
332	tyrT rev	chromosomal tRNA - tyrT rev	gcttctcatccttccccgcat
333	tyrU for	chromosomal tRNA - tyrU for	catggcgcgtcacagttct
334	tyrU rev	chromosomal tRNA - tyrU rev	tctttagacatcgattgtcc
16-1A	rRNA for	chromosomal rRNA intergenic region for	gaatcgctagtaatcg
23-1B	rRNA rev	chromosomal rRNA intergenic region rev	gggttcccccattcgga

CURRICULUM VITA

Elad Firnberg was born in Englewood, New Jersey on October 30, 1984 to Anat and Dov Firnberg. He attended high school at the Academy for the Advancement of Science and Technology of the Bergen County Academies in Hackensack, New Jersey and graduated in 2003. Elad received his undergraduate training and B.E. in chemical engineering from The Cooper Union for the Advancement of Science and Technology in New York City, graduating in May 2008. During this time he spent two summers conducting research in interfacial phenomena at the City College of New York under Dr. Charles Maldarelli and Dr. Alexander Couzis. Elad began his doctoral studies in September 2008 in the Chemical and Biomolecular Engineering Department at Johns Hopkins University in the laboratory of Dr. Marc Ostermeier.